numero 95
luglio 1976
anno IX
volume XVII

LE SCIENZE
edizione italiana di SCIENTIFIC AMERICAN

Il metabolismo dell'alcool

L'eccessivo consumo di bevande alcooliche può causare cirrosi e morte non solo perché favorisce la denutrizione, ma anche perché l'alcool e i suoi prodotti alterano il metabolismo del fegato e ne danneggiano le cellule

di Charles S. Lieber

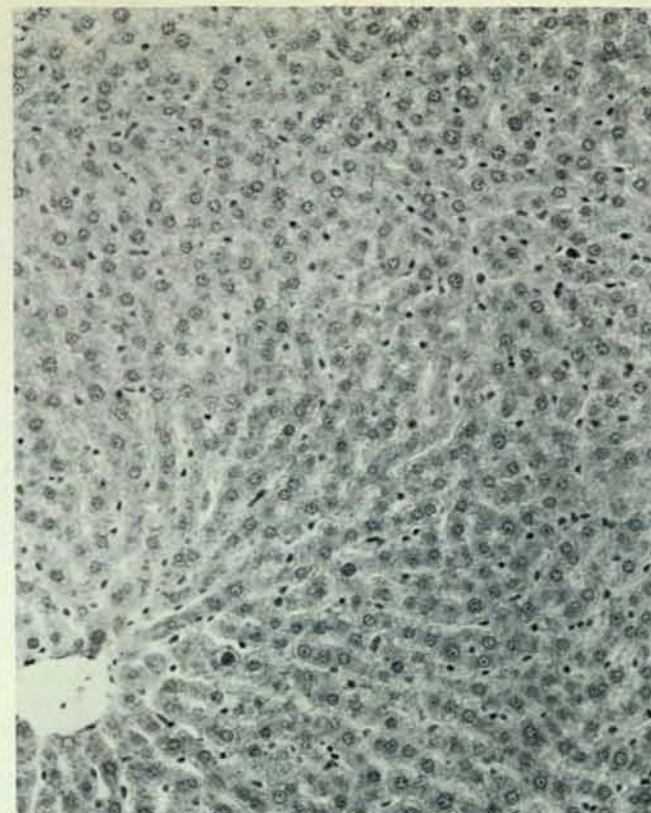
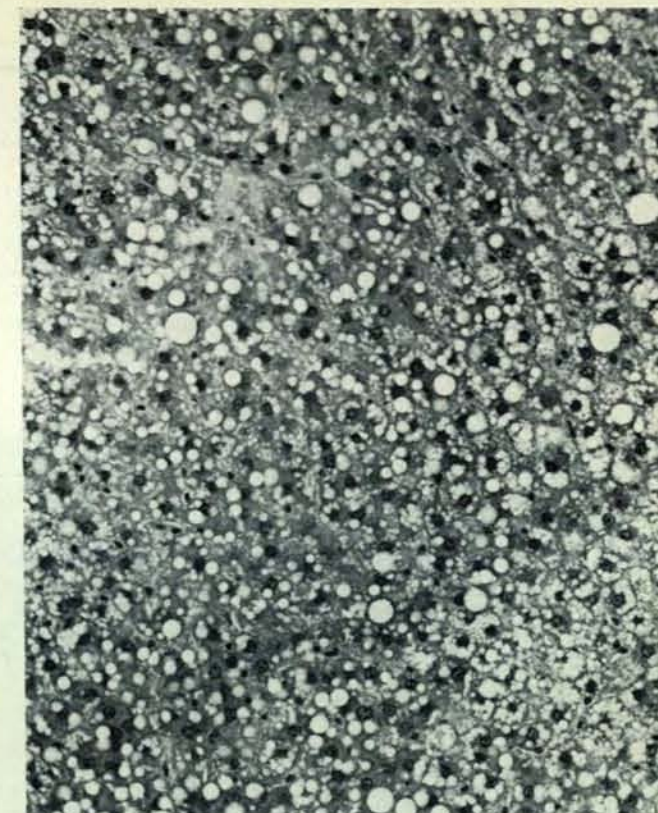
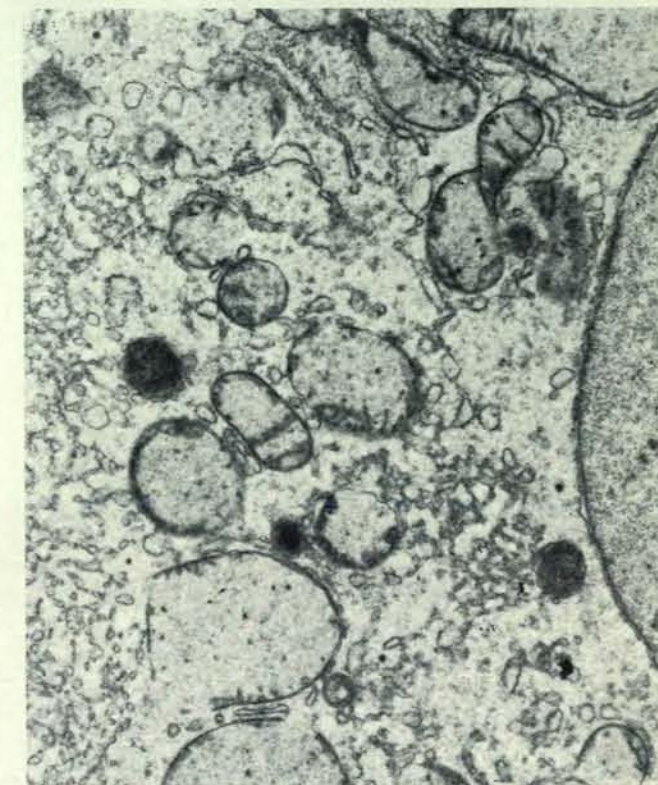
Per quanto l'opinione pubblica si occupi prevalentemente degli effetti dell'eroina, della cocaina e della marijuana, la droga psicotropa più diffusa negli Stati Uniti e in quasi tutte le altre società umane è l'alcool. I suoi effetti psichici, sia positivi che negativi, sono ben noti. Quello che è meno noto è che l'alcool, a dosi diverse secondo gli individui, può rappresentare una vera e propria droga ad azione tossica: il suo eccessivo consumo altera l'intera economia dell'organismo, induce lesioni patologiche nelle cellule del fegato e modifica la funzionalità epatica provocando invalidità e morte. La percentuale degli alcoolizzati fra la popolazione statunitense è andata aumentando parallelamente all'incidenza della cirrosi epatica, la quale nel 1974 ha superato arteriosclerosi, influenza e polmonite sino a raggiungere il settimo posto fra le principali cause di morte; in alcuni centri urbani (inclusa New York) la cirrosi rappresenta oggi, per frequenza, la terza fra le cause di morte dei soggetti in età compresa fra i 25 e i 65 anni. Questa forma morbosa merita uno studio particolare. Soltanto da poco le alterazioni del tessuto epatico e della funzionalità del fegato sono state direttamente correlate a fasi specifiche del metabolismo alcoolico, consentendo così per la prima volta di sperare che si possano elaborare metodi razionali di prevenzione e cura delle affezioni epatiche dovute all'alcool.

Quando si parla di alcool si intende, ovviamente, l'alcool etilico o etanolo (CH_3CH_2OH). L'etanolo è molto probabilmente antico quanto la vita stessa: se per l'uomo è una bevanda, per i lieviti che lo producono costituisce soltanto un prodotto di scarto. L'etanolo acquista le sue caratteristiche soltanto dopo la fermentazione, il processo mediante il quale i lieviti, attraverso l'azione di loro particolari enzimi, ottengono l'energia dai vari zuccheri vegetali. L'uomo ha con molta probabilità conosciuto l'etanolo sin dai tempi preistorici sotto forma di succhi di frutta naturalmente fermentati (vino), miele (idromele), cereali trasformati in malto (birra).

Sin dalle osservazioni anatomiche compiute da Vesalio nel XVI secolo si è riconosciuto che l'eccessiva ingestione di alcool si accompagnava a malattie di parecchi tessuti, soprattutto di quelli del fegato. Sino a poco tempo fa, tuttavia, tali malattie erano attribuite non all'alcool in sé, ma alla denutrizione che spesso si associa all'abitudine di ingerire bevande alcooliche in quantità eccessiva. L'alcool non era considerato una droga, ma un particolare tipo di alimento con determinati effetti psicotropi, il quale poteva essere metabolizzato dall'organismo come un qualsiasi altro nutrimento energetico. Ancora nel 1949 il famoso fisiologo Charles H. Best e i suoi colleghi scrivevano che il contributo metabolico dell'alcool era soltanto quello di fornire calorie e che «non si avevano prove di un effetto tossico specifico dell'alcool etilico puro sulle cellule epatiche più di quante se ne avessero per lo zucchero». Forse era lo stesso apprezzamento per le bevande alcooliche di una parte della popolazione (medici compresi) che dava come fatto accettato il concetto che l'alcool fosse privo di effetti tossici.

Si trattava comunque di una convinzione piuttosto ingenua. L'etanolo ha ben poco in comune con gli altri composti ad alto contenuto energetico. Carboidrati e lipidi possono essere sintetizzati all'interno dell'organismo oppure ingeriti con gli alimenti, mentre l'alcool rappresenta una sostanza essenzialmente estranea al corpo. Come i carboidrati e i lipidi, l'alcool possiede un elevato valore calorico ed è rapidamente assorbito attraverso il tratto gastrointestinale, tuttavia, non può venire immagazzinato nei tessuti. Inoltre, solo quantità molto ridotte di alcool possono essere eliminate attraverso i polmoni e i reni in modo che l'organismo può sbarazzarsene soltanto ossidandolo. Ancora, diversamente da carboidrati e lipidi, che vengono ossidati in quasi tutti i tessuti, l'alcool può essere sottoposto a ossidazione solamente nel fegato, l'organo che contiene la maggior parte degli enzimi necessari a iniziare tale processo. L'organo-specificità dell'alcool spiega i suoi deleteri effetti a livello del fegato, che rappresenta l'area di maggior attività chimica del corpo umano, il centro primario dei processi metabolici che vanno dalla sintesi delle proteine all'azione disintossicante nei confronti dei farmaci e di altre sostanze.

Scopo del presente lavoro è di riferire come le ricerche condotte dall'autore presso il Bronx Veterans Administration Hospital e la Mount Sinai School of Medicine dell'Università di New York dimostrino che la tossicità dell'alcool verso il fegato sia indipendente dalla denutrizione e come fasi specifiche del metabolismo dell'alcool siano correlate ad alterazioni del fegato e di altri tessuti.

Esistono motivi socioeconomici e fisiologici che spiegano la denutrizione nel soggetto alcoolizzato. Questi infatti spreca tempo, denaro ed energie nel bere

Degenerazione grassa del fegato (o steatosi) ottenuta in ratti di laboratorio mediante somministrazione di alcool e malgrado una dieta sufficiente. In queste microfotografie sono visibili sezioni di tessuto epatico di ratto ingrandite 240 volte. A sinistra, tessuto di ratto mantenuto a dieta liquida priva di alcool. A destra, tessuto epatico di ratto a cui è stata somministrata una dieta liquida contenente alcool pari al 36 per cento delle calorie totali. Dopo 24 giorni la microfotografia del fegato del secondo ratto mostra la presenza di numerose goccioline lipidiche.

Alterazioni microstrutturali rivelate da microfotografie elettroniche eseguite da Oscar A. Iseri e dall'autore in cui una porzione della cellula epatica è ingrandita 16 mila volte. Nel fegato di controllo (*a sinistra*) gli organelli grigi con membrane introflesse (creste) sono mitocondri normali; fra di essi sono visibili strie parallele di reticolo endoplasmico rugoso costellato di ribosomi unitamente a reticolo endoplasmico liscio (sacciforme). Nel fegato del ratto cui è stato somministrato etanolo (*a destra*) i mitocondri appaiono rigonfiati e deformati; in alcuni la membrana esterna e le creste sono distrutte. Si noti una marcata proliferazione del reticolo endoplasmico liscio.

e può trascurare la preparazione degli alimenti e dimenticare anche di mangiare. Diversamente dalle altre droghe, l'alcool possiede un elevato valore calorico: 7,1 calorie per grammo, cosicché mezzo litro di una bevanda alcoolica a circa 40 gradi (consumo non insolito per un alcoolizzato) rappresenta pressappoco la metà del fabbisogno quotidiano calorico. Di conseguenza può diminuire il desiderio per il cibo. Le calorie dell'etanolo sono, tuttavia, «calorie inutilizzabili» che non forniscono proteine, minerali e vitamine. Inoltre, l'alcool può aggravare direttamente la denutrizione. Provocando gastrite, pancreatite e colite può alterare la digestione e l'assorbimento; la denutrizione può, a sua volta, diminuire la funzionalità intestinale. Infine, l'etanolo e il suo principale prodotto di conversione, l'acetaldeide, possono interferire con l'attivazione delle vitamine da parte delle cellule epatiche. Di conseguenza, la denutrizione è un fenomeno comunemente osservabile negli alcoolizzati ed essa, da sola, può alterare la funzionalità epatica, come hanno dimostrato in maniera inequivocabile le osservazioni eseguite sugli animali da esperimento mantenuti a diete gravemente carenti.

Si dimostra dunque opportuno verificare quanto incida la denutrizione sulle malattie epatiche dovute al consumo di alcool anche se si ammette che l'alcool non è un alimento. Resta il fatto che nella pratica medica si incontrano alcoolizzati con malattie epatiche malgrado l'osservanza di una dieta apparentemente adeguata. Sul finire degli anni cinquanta l'autore del presente lavoro ha incominciato a chiedersi se l'alcool fosse in grado di esercitare un'azione tossica diretta sul fegato. Una tale osservazione avrebbe avuto un rilevante interesse sia terapeutico sia teorico. Molti medici a quei tempi dicevano ai loro pazienti alcoolizzati che una dieta adeguata avrebbe potuto conservare una normale funzionalità epatica malgrado il consumo di bevande alcooliche.

L'importanza del problema e la convinzione che anche un elevato consumo di alcool provocasse un rischio tollerabile, se effettuato per breve periodo e sotto un rigoroso controllo, ha spinto l'autore a studiare direttamente l'effetto dell'etanolo in soggetti volontari. Questi sono stati mantenuti a una dieta ottimale con bassa percentuale di lipidi; in tale dieta le proteine rappresentavano il 25 per cento delle calorie ossia due volte e mezzo la percentuale raccomandata. Ciò significava carne di manzo alla prima colazione, formaggio e pesce a mezzogiorno e carne o pollame alla sera, con una somministrazione supplementare di minerali e vitamine. I volontari bevevano anche sei volte al giorno, per un totale di 283 grammi di alcool a circa 40 gradi. Questo regime alimentare ha indotto un progressivo aumento del grasso epatico; la biopsia di controllo rivelava dopo alcuni giorni un significativo incremento del contenuto lipidico nel fegato e dopo 18 giorni di tale esperimento l'incremento medio era di otto volte. Si osservava la caratteristica degenerazione grassa del fegato o steatosi, il primo stadio (tipicamente reversibile) della malattia epatica. Apparivano anche marcate alterazioni nella struttura delle cellule epatiche: i mitocondri, gli organelli deputati a trasformare l'energia, risultavano allargati e deformati mentre si registrava una proliferazione delle membrane del reticolo endoplasmico, il sito degli enzimi associati al metabolismo dell'alcool e delle altre sostanze. Queste modificazioni del fegato erano provocate da un'ingestione piuttosto moderata di etanolo, tale da non indurre segni clinici di intossicazione; i valori ematici oscillavano da 80 a 90 milligrammi di alcool per 100 millilitri di sangue, al di sotto cioè dei valori (100 o 150 milligrammi) che rappresentano la prova di un'eccessiva libagione in molti paesi.

Per verificare questi effetti e studiare in maniera particolareggiata il loro meccanismo si doveva riprodurli negli animali sperimentali. Di solito il ratto di laboratorio costituisce il soggetto di tali esperimenti. L'etanolo viene solitamente somministrato a questi animali mediante

Cirrosi epatica indotta in un babbuino al quale sono stati somministrati per quattro anni consecutivi elevati quantitativi di alcool etilico. Questa sezione di fegato, ingrandita 70 volte, mostra la presenza di spessi filamenti di tessuto connettivo fibroso, simile a tessuto cicatriziale, che hanno alterato la precedente regolare struttura delle cellule epatiche, separando formazioni nodulari costituite da cellule irregolarmente disposte e densi ammassi di goccioline lipidiche.

Differenti stadi dell'ossidazione dell'alcool etilico (etanolo). Nelle cellule epatiche due atomi di idrogeno vengono staccati da ciascuna molecola di etanolo per formare acetaldeide. Questa è di norma ossidata principalmente nel fegato per sintetizzare acido acetico (sotto forma di acetati), poi alla fine trasformato in anidride carbonica e acqua. Nell'illustrazione sono tralasciati i vari enzimi e i cofattori.

Rappresentazione schematica del metabolismo dell'etanolo nelle cellule epatiche mediante il sistema dell'alcooldeidrogenasi (ADH) e quello di ossidazione dell'etanolo microsomale (MEOS), con le due vie dell'etanolo e i movimenti dei loro prodotti indicati in colore. Le vie che operano in assenza dell'etanolo sono indicate in nero e quelle la cui attività è diminuita dall'etanolo sono evidenziate dalle linee tratteggiate. Nella via primaria dell'etanolo l'alcooldeidrogenasi catalizza lo spostamento degli atomi di idrogeno (H) mediante riduzione di un cofattore, la nicotinammideadenindinucleotide (NAD). L'idrogeno in eccesso è smistato in vari processi. Per esempio, rifornisce il sistema energetico di trasporto degli elettroni sostituendo i grassi. L'idrogeno contribuisce anche alla sintesi dell'eccesso di trigliceridi. Una via metabolica secondaria entra in azione ad alti livelli alcoolici utilizzando il sistema microsomale che metabolizza anche alcuni farmaci e altri composti estranei all'organismo. In questo caso il cofattore è nicotinammideadenindinucleotidefosfato ridotto (NADPH).

l'acqua che essi bevono, tuttavia, e in queste condizioni, gli animali in genere rifiutano di bere una quantità di alcool sufficiente a provocare una lesione epatica. Leonore M. De Carli e l'autore hanno superato l'avversione dei ratti per l'alcool sviluppando una nuova tecnica: l'etanolo veniva introdotto in una dieta nutritivamente adeguata, ma totalmente liquida in modo che gli animali, per mangiare o bere, dovevano forzatamente assumere anche l'alcool. Con questo metodo è stato possibile indurre la comparsa di una degenerazione grassa del fegato malgrado la presenza di una dieta adeguata e totalmente controllata. Le lesioni nel fegato erano perfettamente comparabili a quelle rilevabili nell'uomo sia a un esame macroscopico, sia al controllo istologico effettuato mediante microscopio ottico (*si vedano le illustrazioni in alto a pagina 9*). Anche le alterazioni ultrastrutturali, osservate con il microscopio elettronico, apparivano similari, includendo dilatazione e distruzione dei mitocondri e proliferazione del reticolo endoplasmico, (*si vedano le illustrazioni in basso a pagina 9*).

Per quanto i ratti trattati con l'alcool sviluppassero una degenerazione grassa del fegato, non manifestarono gli stadi più gravi di lesione epatica rilevabili nei soggetti umani. Il primo di questi stadi è costituito dall'epatite alcoolica in cui la ridotta funzionalità cellulare del fegato porta a necrosi cellulare, processi infiammatori e morte con una frequenza del 10-30 per cento. Lo stadio finale è rappresentato dalla cirrosi, nella quale le cicatrizzazioni fibrose distruggono la normale architettura del fegato e danno inizio ad alcune complicazioni potenzialmente mortali. Si potrebbero ipotizzare due motivi che preservano i ratti dall'epatite e dàlla cirrosi: anche con la dieta liquida utilizzata, la loro ingestione di alcool non superava il 36 per cento delle calorie totali, il che corrisponde a un consumo moderato per l'uomo. Inoltre, per l'uomo occorre un periodo di 5-25 anni di costante consumo di bevande alcooliche per sviluppare una cirrosi e il ratto vive, di regola, circa due anni.

Il grasso epatico deriva da tre fonti: dalla dieta, attraverso l'intestino; sotto forma di acidi grassi liberi mobilizzati dai depositi del tessuto adiposo e mediante una sintesi che si verifica all'interno del fegato stesso. L'insorgenza di una degenerazione grassa del fegato può essere legata sia a un'eccessiva disponibilità di grassi, sia a un'alterazione nella loro normale eliminazione.

L'autore ha quindi utilizzato i babbuini, che vivono a lungo e sono strettamente correlati all'uomo, tanto da essere capaci di tollerare diete liquide contenenti alcool pari al 50 per cento delle calorie totali, corrispondente circa a quanto assume un essere umano alcoolizzato. In seguito a questo trattamento i babbuini apparivano visibilmente intossicati e quando la somministrazione di alcool veniva temporaneamente sospesa, manifestavano segni di dipendenza fisica, come tremori e manifestazioni accessuali. Presso il laboratorio di medicina sperimentale dei primati, l'autore ha mantenuto 16 babbuini a una dieta ricca in etanolo e 16 altri a una dieta priva di alcool ma con lo stesso contenuto calorico. Mentre negli animali di quest'ultimo gruppo il fegato si è conservato normale, negli altri animali si è presto registrato un eccessivo accumulo di grasso epatico; cinque di loro hanno inoltre accusato una tipica epatite alcoolica e in sei è comparsa una cirrosi dopo un periodo di due-quattro anni. Gli esperimenti condotti sui ratti e sui babbuini hanno così dimostrato in maniera evidente che una prolungata ed elevata ingestione di alcool può indurre gravi lesioni epatiche anche se la dieta viene mantenuta a un livello normale. Si è in questo modo anche potuto ottenere per la prima volta un modello animale sperimentale in grado di riprodurre tutte le lesioni epatiche rilevabili negli esseri umani. Questi modelli hanno anche consentito di trovare spiegazioni biochimiche per gli stadi precoci della malattia alcoolica epatica (degenerazione grassa del fegato e turbe metaboliche associate) e per alcune delle lesioni croniche che sembrano portare agli stadi più gravi: epatite e cirrosi.

La prima fase della via metabolica primaria dell'alcool è catalizzata da un enzima, la deidrogenasi alcoolica (malgrado il suo nome, la deidrogenasi alcoolica trasferisce atomi di idrogeno da vari composti, inclusi alcuni steroidi; era pertanto già presente nel fegato dell'uomo preistorico che per primo sperimentò l'alcool). L'alcooldeidrogenasi catalizza il trasporto degli atomi di idrogeno dall'etanolo al cofattore, nocotinammideadenindinucleotide (NAD), che trasforma l'etanolo in acetaldeide. Quest'ultima è poi ossidata, principalmente nel fegato, per formare acetato che alla fine del processo è convertito in anidride carbonica e acqua. Alcuni effetti metabolici dell'alcool sono direttamente connessi ai due primi prodotti di ossidazione: idrogeno e acetaldeide.

L'eccesso di idrogeno proveniente dall'alcool sbilancia l'attività chimica delle cellule epatiche. Per sopravvivere, la cellula deve liberarsi dell'idrogeno e lo fa smistando gli ioni idrogeno in uno o più vie metaboliche dipendenti, talvolta con effetti deleteri. Una di tali vie metaboliche è rappresentata dal processo per cui gli amminoacidi (derivati dalla dissociazione delle proteine nel fegato) sono trasformati in glucosio, con formazione di piruvato come preparato intermedio. In presenza di un eccesso di ioni idrogeno il processo segue una diversa direzione: il piruvato è ridotto a lattato invece di trasformarsi in glucosio. Lo zucchero ematico proviene da tre fonti: gluconeogenesi, o sintesi dagli amminoacidi nel fegato, dissociazione del glicogeno immagazzinato nei tessuti e trasformazione dei carboidrati contenuti nella dieta. Se l'individuo alcoolizzato beve e non mangia, non ingerisce carboidrati alimentari e consuma le proprie riserve di glicogeno; se la glucogenesi è poi bloccata dalla trasformazione del piruvato in lattato, risulterà abbassata la glicemia. Il basso contenuto di zucchero nel sangue, o ipoglicemia, rappresenta una nota complicazione dell'alcoolismo acuto, ma è spesso trascurata. Quando un individuo intossicato è ricoverato in un reparto di pronto soccorso è importante accertare l'eventuale presenza di una ipoglicemia, giacché organi di importanza vitale, compreso il cervello, possono essere colpiti in maniera critica da una carenza di zucchero; alcuni casi di morte osservabili in alcoolizzati possono essere dovuti a questa condizione.

Il lattato che si forma a causa di un eccesso di idrogeno presenta conseguenze diverse: entra nel sangue inducendo una acidosi lattica e, a livello dei reni, interferisce con l'escrezione di acido urico. Un elevato contenuto di acido urico nel sangue (iperuricemia) aggrava la gotta cosicché il processo qui descritto può spiegare le antiche osservazioni cliniche secondo le quali eccessive libagioni potevano scatenare o aggravare la malattia.

Esistono altri modi attraverso i quali la cellula epatica si libera da un eccesso di idrogeno. Alcuni di questi implicano la formazione di lipidi, o grassi. L'idrogeno può essere portato direttamente nella sintesi dell'alfaglicerofosfato e degli acidi grassi. Questi sono i due precursori dei trigliceridi e i trigliceridi sono, a loro volta, i lipidi che si accumulano in caso di degenerazione grassa del fegato a eziologia alcoolica. Il principale meccanismo per eliminare l'idrogeno è più indiretto, ma il risultato è simile. L'idrogeno è trasportato nei mitocondri, gli organelli cellulari che producono l'energia necessaria per le funzioni epatiche. In condizioni normali è il grasso a essere ossidato - in effetti bruciato - nel ciclo dell'acido citrico mitocondriale per produrre l'energia utilizzabile sotto forma di ioni fosfato altamente energetici. L'idrogeno abbondantemente fornito dall'alcool, tuttavia, garantisce una sostanza alternativa che è ossidata al posto dell'idrogeno proveniente dai grassi. Tale sostituzione esige però un prezzo; l'accumulo di lipidi, che portano alla degenerazione grassa del fegato. Se l'alcool è

Nei ratti a cui viene somministrato alcool si manifesta una tendenza all'iperlipemia, cioè a un elevato contenuto di grassi nel sangue. Gli animali per un mese sono stati tenuti sia a una dieta di controllo sia a una dieta contenente alcool con lo stesso valore calorico. Poi, dopo un periodo di digiuno, agli animali è stato dato un «carico» costituito da una o dall'altra dieta. Il contenuto dei lipidi nel sangue era pressappoco lo stesso nei ratti a digiuno. In risposta al carico gli animali consumatori cronici di alcool (*a destra*) hanno manifestato una lipemia più elevata sia che il carico contenesse alcool (*barre in colore*) sia che ne fosse sprovvisto (*barre in grigio*).

L'alcool attiva i sistemi enzimatici del reticolo endoplasmico liscio aumentando il metabolismo di talune sostanze. Gli andamenti illustrano la scomparsa dal sangue di un tranquillante, il meprobamato, somministrato a volontari prima (*in nero*) e dopo (*in colore*) un mese di consumo di alcool. L'attivazione alcoolica taglia a metà (*linee tratteggiate*) la semivita del farmaco.

Le calorie alcooliche non sembrano del tutto equivalenti alle altre calorie. Quando una dieta base di 2200 calorie viene completata da 2000 calorie fornite da alcool (*a sinistra*) o da cioccolato (*a destra*) l'incremento del peso corporeo risulta molto meno marcato nel primo caso.

ingerito insieme a una dieta contenente grassi, questi ultimi si accumulano nel fegato; ma anche quando l'alcool è consumato contemporaneamente a una dieta a basso contenuto lipidico si ha un deposito nel fegato del grasso che vi viene sintetizzato. Inoltre, quando l'alcool viene bevuto in grandi quantitativi, può scatenare la messa in circolo di ormoni che mobilizzano i grassi immagazzinati nel tessuto adiposo e li indirizzano verso il fegato (*si veda l'illustrazione a pagina 12*).

Cosa ne fa il fegato del grasso accumulato? Esso può essere secreto nella massa sanguigna che fornisce le sostanze utili per i tessuti periferici, quali i muscoli, e deposita quelle in eccesso nel tessuto adiposo. La secrezione è complicata dal fatto che i lipidi sono insolubili nell'acqua; essi devono esser resi solubili mediante la copertura con un sottile rivestimento di proteine per formare le lipoproteine. L'assemblaggio di queste lipoproteine viene effettuato nel fegato, a livello delle membrane del reticolo endoplasmico. È stata ricordata la proliferazione del reticolo endoplasmico che avviene in seguito a un elevato consumo di alcool sia nei ratti sia negli esseri umani: tale fenomeno è stato osservato dall'autore insieme a Oscar A. Iseri, Bernard P. Lane ed Emanuel Rubin. Tale proliferazione si rispecchia, come hanno riscontrato Jean-Gil Joly, Lawrence Feinman e l'autore, nell'esaltata attività di certi enzimi presenti nel reticolo endoplasmico capaci di aumentare l'idoneità del fegato a secernere lipoproteine. Un fegato adattatosi all'alcool in seguito a un breve condizionamento provocato dal consumo di elevate quantità di alcool, reagirà con un'eccessiva secrezione di lipoproteine anche dopo un pasto normale, producendo una iperlipemia, ossia un livello anormalmente elevato di grassi nel sangue. Enrique Baraona e l'autore hanno riscontrato questo effetto nei ratti mentre William H. Perlow, Stephen A. Borowsky e lo stesso autore lo hanno ora verificato nell'uomo. Questo effetto è di particolare importanza per quei soggetti che presentano anomalie sia del metabolismo dei lipidi sia di quello dei carboidrati e sono pertanto inclini a elevati contenuti di lipidi nel sangue. L'iperlipemia rappresenta uno dei maggiori fattori di rischio per gli attacchi cardiaci. Quello che di solito viene trascurato è che in individui con una preesistente iperlipemia l'alcool probabilmente è l'unico fattore aggravante suscettibile di correzione.

Un'altra via metabolica a disposizione del fegato per diminuire l'eccesso di grassi è quella di convertirne una parte nei prodotti di degradazione solubili in acqua, chiamati corpi chetonici, e immetterli nella circolazione sanguigna. In qualche soggetto particolarmente sensibile, questa risposta può essere esagerata: si ha in tal caso un incremento dei corpi chetonici nel sangue che produce quella condizione nota come acidosi nei pazienti diabetici.

La trasformazione dei grassi in lipoproteine è soltanto una delle numerose funzioni del reticolo endoplasmico della cellula epatica, che inattiva anche una grande varietà di farmaci e altre sostanze estranee, convertendole in prodotti idrosolubili che possono essere escreti (si veda l'articolo *Come il fegato metabolizza le sostanze estranee all'organismo,* di Attalah Kappas e Alvito P. Alvares, in «Le Scienze», n. 85, ottobre 1975). L'autore si è pertanto chiesto se la proliferazione del reticolo endoplasmico, dopo un consumo cronico di alcool, si sarebbe riflessa in un'aumentata capacità del fegato dell'alcoolizzato a metabolizzare i vari farmaci. Dopo ripetute somministrazioni di alcool (che però non avevano ancora dato origine a serie lesioni al fegato), gli enzimi del reticolo endoplasmico che inattivano i tranquillanti, gli anticoagulanti e altri farmaci in grado di disintossicare l'organismo da certi additivi alimentari, sostanze cancerogene e insetticidi, aumentano la loro attività accrescendo la capacità dell'organismo di liberarsi da tali composti. Per esempio, Prem S. Misra e l'autore hanno somministrato un tranquillante, il meprobamato, misurando la sua curva di decadimento nel sangue. Dopo un mese di ingestione di alcool, il tempo necessario perché il livello ematico del farmaco diminuisca a metà del suo valore si riduce da sedici a otto ore.

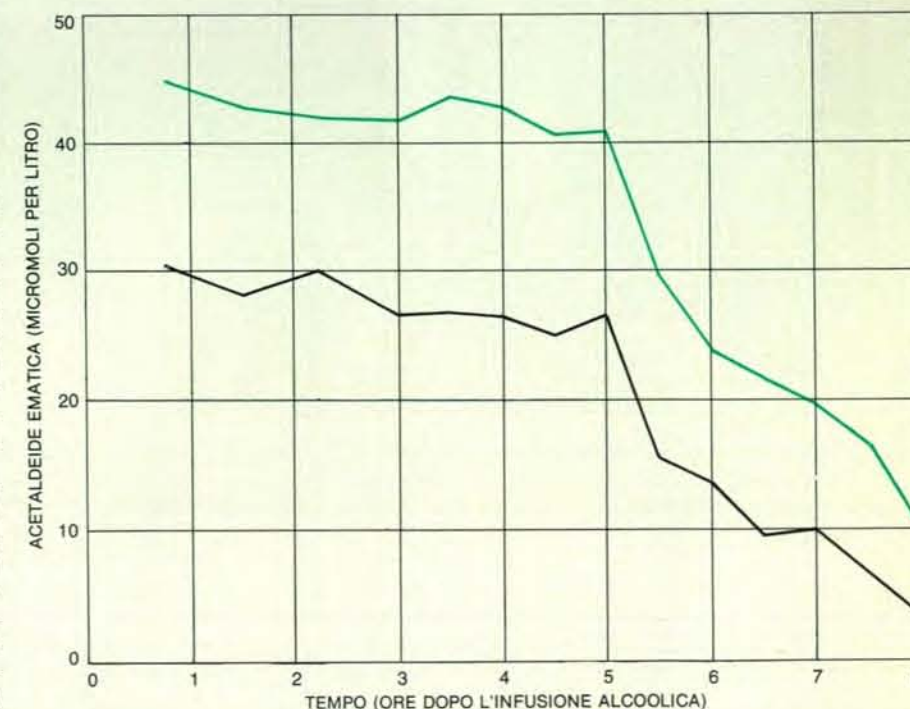
Gli anestesiologi sanno da molti anni che per raggiungere l'effetto desiderato negli alcoolizzati sono richieste dosi di sedativi più elevate che negli altri soggetti. Questa «tolleranza» ai farmaci è stata attribuita a una forma di adattamento da parte del sistema nervoso centrale: il cervello degli alcoolizzati avrebbe cioè una aumentata resistenza ai sedativi. La ricerca degli autori ha messo invece in rilievo un adattamento metabolico che faceva aumentare la capacità del fegato dell'alcoolizzato a inattivare ed eliminare sedativi e altri composti che il reticolo endoplasmico ha la funzione di disintossicare. A uno stadio non troppo avanzato della loro malattia, gli alcoolizzati richiedono perciò maggiori dosi di molti farmaci. Questo è vero, tuttavia, soltanto quando l'individuo è sobrio. Quando l'alcoolizzato ha bevuto, l'effetto è pressappoco l'opposto. Il motivo, è che uno dei farmaci che il reticolo endoplasmico metabolizza, è l'alcool stesso, tramite una via accessoria che coadiuva il sistema alcooldeidrogenasi di base.

De Carli e l'autore hanno dimostrato l'esistenza di una via accessoria ultracentrifugando tessuto epatico e isolando il reticolo endoplasmico, detto anche frazione microsomale. Si scoprì che un preparato della frazione microsomale era in grado di ossidare l'alcool etilico. Questa via accessoria fu denominata sistema microsomale alcool etilico-ossidante e Rolf Teschke, Kunihiko Ohnishi e l'autore sono stati in grado di ottenerlo in una forma semipurificata. Si è anche potuto appurare che questa via microsomale per l'alcool entra in funzione dopo che il livello alcoolico ematico raggiunge un certo livello. L'alcool entra perciò in competizione con altri farmaci, al cui metabolismo partecipano vari elementi del sistema microsomale, ritardandone pertanto il metabolismo e aumentandone di conseguenza l'effetto. Questa è la causa per cui è particolarmente pericoloso bere e contemporaneamente prendere tranquillanti: l'alcool può accentuare l'azione del farmaco non solo perché l'effetto dei due farmaci sul cervello può sommarsi, ma anche perché la presenza di alcool può interferire con la capacità epatica di inattivare il farmaco, cosicché a un certo dosaggio la maggior parte del medicamento rimane attiva per un tempo maggiore.

Con il supporto dei sistemi enzimatici microsomali, il sistema alcool etilico-ossidante adatta se stesso all'eccessivo consumo di alcool, aumentando la propria attività e contribuendo così alla «tolleranza» per l'alcool. La capacità dell'alcoolizzato di bere in misura maggiore rispetto alla maggioranza dei non alcoolizzati è principalmente dovuta alla capacità di resistenza del cervello e a una progressiva diminuzione dell'effetto cerebrale dell'alcool. L'alcoolizzato sviluppa inoltre un'aumentata capacità di metabolizzare alcool non solo attraverso il sistema microsomale, ma anche tramite l'alcooldeidrogenasi e, forse, anche attraverso una terza via che dipende dall'enzima catalasi. Dopo eccessive o prolungate libagioni di bevande alcooliche questo adattamento, tuttavia, può essere controbilanciato da progressive lesioni epatiche di modo che la capacità globale del fegato di metabolizzare l'alcool rimane circa la stessa oppure addirittura diminuisce.

Le modificazioni microsomali descritte dall'autore sono positive nel senso che aiutano il fegato a liberarsi dai grassi e accelerano la disintossicazione di molti farmaci, additivi alimentari e altri composti estranei. L'iperlipemia è una indesiderata concomitanza del processo adattativo, ma ne esistono anche altre. Alcune sostanze estranee all'organismo sono attivate, piuttosto che inattivate, dalle trasformazioni che esse subiscono nel reticolo endoplasmico. Alcune sostanze potenzialmente cancerogene diventano tali solo dopo attivazione da parte dei microsomi e altre sostanze divengono tossiche per il fegato stesso solo dopo tale attivazione. Per esempio, l'esposizione al tetracloruro di carbonio può causare danni epatici, ma il familiare composto per pulire a secco è innocuo per il fegato fino a che non sia attivato dal reticolo endoplasmico della cellula epatica. Yasushi Hasumura e l'autore hanno osservato che l'incremento dell'attività microsomale indotta dall'alcool aumenta la tossicità del tetracloruro di carbonio: ratti trattati per un periodo prolungato con l'alcool apparivano molto più sensibili agli effetti tossici del suddetto composto degli animali di controllo. Questo effetto spiega presumibilmente l'osservazione clinica secondo la quale gli alcoolizzati sono particolarmente sensibili all'avvelenamento da tetracloruro di carbonio negli impianti per la pulitura a secco che ancora usano questo composto. L'ipersensibilità degli alcoolizzati molto probabilmente si estende a un gran numero di sostanze estranee che sono del tutto innocue per la maggior parte della gente.

Un altro effetto secondario consiste in uno spreco energetico. L'attività microsomale richiede infatti energia; oltre tutto è peculiare delle varie ossidazioni microsomali produrre calore senza conservare energia chimica. Questo fenomeno potrebbe essere la causa della diminuita crescita osservata negli animali nutriti con etanolo dal momento che la produzione del calore, oltre quello richiesto per mantenere la temperatura corporea, costituisce uno spreco di energia. Tale spreco può, almeno in parte, spiegare l'osservazione compiuta dall'autore unitamente a Romano C. Pirola, che l'aggiunta di alcool etilico alla dieta produce un aumento del peso corporeo inferiore a quello ottenuto con uno stesso numero di calorie provenienti però da altre fonti (*si veda l'illustrazione nella pagina a fronte*).

A prescindere dalle complicazioni metaboliche di un eccesso di idrogeno e dalle variazioni dell'attività microsomale, la forte ingestione di alcool esplica effetti tossici diretti sul tessuto epatico. Sembra che una funzione particolare, per quanto si riferisce alla tossicità dell'alcool, sia svolta dall'acetaldeide, un prodotto delle numerose vie metaboliche dell'alcool, estremamente attiva e capace di intaccare molti tessuti dell'organismo. Molta parte dell'acetaldeide è trasformata in acido acetico dai mitocondri epatici, ma una parte finisce nel sangue. L'alcooldeidrogenasi e le vie microsomali alcool etilico-ossidanti si saturano entrambe quando il fegato è investito da una quantità rilevante di alcool, cosicché il livello di acetaldeide nel sangue raggiunge un plateau e rimane a tale livello finché il valore dell'alcool diminuisce fino al punto in cui apparentemente interseca la via microsomale. Mark A. Korsten, Shohei Matsuzaki, Feinman e l'autore hanno potuto scoprire che il plateau dell'acetaldeide risulta significativamente più elevato negli alcoolizzati che nei non alcoolizzati, anche quando a entrambi i gruppi viene dato lo stesso quantitativo di alcool ed è stato raggiunto lo stesso livello di alcool nel sangue (*si veda la figura qui in basso*). Questo livello di ace-

Il contenuto ematico in acetaldeide è più elevato nei soggetti alcoolizzati che in quelli normali. Dosi analoghe di etanolo puro sono state iniettate nel sangue di soggetti alcoolizzati e non alcoolizzati sino a raggiungere uguali valori plasmatici. Il *plateau* dell'acetaldeide è risultato più elevato negli alcoolizzati (*in colore*) che negli altri (*in nero*). Sembrerebbe che i primi metabolizzino l'acetaldeide meno efficacemente, forse a causa di lesioni epatiche indotte dall'etanolo.

taldeide, abnormemente elevato negli alcoolizzati, può essere il risultato del più veloce metabolismo dell'alcool nel dare acetaldeide (a causa dell'incremento adattativo dell'attività microsomale), oppure può essere la conseguenza del deterioramento metabolico dell'acetaldeide.

Probabilmente sono implicati entrambi i fattori, almeno all'inizio. Si è in precedenza accennato alle singolari alterazioni nei mitocondri rilevate dal microscopio elettronico anche negli stadi più precoci di forte consumo di alcool. Hasumura e l'autore hanno isolato i mitocondri danneggiati e hanno scoperto una riduzione della loro capacità di metabolizzare l'acetaldeide ad acido acetico. La acetaldeide stessa può essere responsabile di parte del decremento della funzione mitocondriale: Arthur I. Cederbaum, Rubin e l'autore hanno dimostrato che essa ha un effetto tossico sugli organelli. L'alcoolizzato può essere quindi vittima di un circolo vizioso: un alto livello di acetaldeide danneggia la funzionalità mitocondriale nel fegato, il metabolismo dell'acetaldeide diminuisce, la maggior parte di essa si accumula e causa ulteriori danni al fegato. L'acetaldeide danneggia non solo il fegato, ma anche altri tessuti. Per esempio, Marcus A. Rothschild e i suoi collaboratori del Manhattan Veterans Administration Hospital, hanno riferito che livelli di acetaldeide non più elevati di quelli che erano stati misurati nel sangue degli alcoolizzati possono inibire la sintesi proteica nel muscolo cardiaco. Tale effetto può in parte chiarire la diminuita funzionalità cardiaca comune negli alcoolizzati.

L'acetaldeide esplica marcati effetti sul cervello. Molti ricercatori hanno ipotizzato che tale sostanza, piuttosto che l'alcool stesso, sia responsabile dello sviluppo della dipendenza che, assieme alla maggior tolleranza, caratterizza gli alcoolizzati. La dipendenza si manifesta sotto forma di uno stato di estremo sconforto, spesso associato a disturbi fisiologici come tremori e attacchi, quando viene a mancare l'alcool. Molti meccanismi mediati dall'acetaldeide sono stati ipotizzati per spiegare la dipendenza, ma nessuno è stato ancora confermato. Un'ipotesi si basa sul fatto che certe ammine neuromediatrici, le quali trasmettono l'impulso nervoso da una cellula all'altra nel cervello, sono inattivate dalle MAO (monoamminoossidasi) per formare un'aldeide che è poi trasformata in un acido. La trasformazione ad acido richiede un enzima che è attivo anche nel metabolismo dell'acetaldeide. Virginia E. Davis e Michael J. Walsh, dell'Houston Veterans Administration Hospital, hanno ipotizzato che se l'acetaldeide è presente nel cervello essa può competere con l'enzima: le aldeidi neuromediatrici non metabolizzate possono così accumularsi; in tal modo si possono allora combinare con il neuromediatore formando composti molto simili a certi morfinoderivati, che inducono dipendenza.

Un'altra possibilità, proposta da Gerald Cohen del Mount Sinai, è che l'acetaldeide si combini direttamente con le ammine per formare i derivati della isochinolina, potenti composti psicoattivi che possono svolgere una determinata funzione nell'indurre dipendenza. La dipendenza dall'alcool è probabilmente determinata da diversi fattori che agiscono contemporaneamente. È possibile, tuttavia, che l'acetaldeide sia implicata nella predisposizione all'alcoolismo: una persona predisposta può avere deficienze nel metabolismo dell'acetaldeide che intervengono nel mantenere un più elevato livello ematico dell'acetaldeide con le conseguenze che da esso derivano.

Se continua il forte consumo di alcool, la degenerazione grassa del fegato ancora reversibile evolve, nella maggioranza dei casi, verso più gravi e irreversibili malattie epatiche: epatiti e poi cirrosi. Non è stato ancora chiarito il motivo e il modo con cui il fegato perde la sua capacità di adattarsi al carico di alcool. Anche allo stadio della degenerazione grassa esistono alcune indicazioni che si stanno sviluppando lesioni più gravi. Esse possono consistere in un rigonfiamento delle cellule epatiche comunemente attribuito all'accumulo di sostanze lipidiche. Baraona, Maria-Anna Leo, Borowsky e l'autore hanno notato che la capacità del fegato di liberare proteine, di cui questo organo è il maggiore produttore anche quando sono accumulate nelle cellule, è in qualche modo depressa da rilevanti quantitativi di alcool. L'intasamento delle cellule epatiche con lipidi e proteine interferisce con il loro normale funzionamento. Così si verifica la ridotta produzione di energia a cui si era in precedenza accennato. Per tutti questi motivi alcune cellule del fegato possono morire e la necrosi può innescare un processo infiammatorio, caratteristico di quello stadio chiamato epatite cronica.

L'acuta riduzione della funzionalità della cellula epatica, rilevabile a questo stadio, finisce con il provocare la morte in alcuni pazienti.

La necrosi cellulare e l'infiammazione promuovono a loro volta lo stadio successivo: la fibrosi, o sviluppo di tessuto cicatriziale, l'anticamera della cirrosi. Le barriere di tessuto connettivo fibroso fra i gruppi di cellule epatiche interferiscono con l'irrorazione sanguigna compromettendo ulteriormente la funzionalità epatica. Parzialmente bloccato nel fegato, il sangue si ingorga aumentando la pressione del sistema portale, che trasporta il sangue dall'intestino al fegato. Circoli collaterali anomali possono quindi svilupparsi nel sistema venoso, cosicché parte del sangue può «bypassare» il blocco circolatorio provocato dal fegato cirrotico. Nell'esofago possono comparire formazioni venose di tipo varicoso, suscettibili di rottura ed emorragia: le varici sanguinanti sono una delle maggiori cause di morte nella cirrosi. Inoltre, in conseguenza dell'alta pressione, il plasma filtra attraverso i vasi sanguigni del sistema portale. Si forma così linfa che fuoriesce dai vasi linfatici. Entrambi questi fattori contribuiscono alla formazione delle asciti, ossia all'accumularsi di liquido nella cavità addominale. Una complicazione finale è dovuta all'incapacità del sangue di liberarsi dall'ammoniaca presente nel sangue e da altri composti azotati, prodotti dai batteri intestinali. Appena questi composti hanno raggiunto un certo valore, agiscono sul cervello e possono provocare turbe funzionali, coma epatico e morte.

Complicazioni da eccessivo consumo di alcool derivano prevalentemente da un eccesso di idrogeno e dall'acetaldeide. L'idrogeno produce degenerazione grassa del fegato (steatosi), iperlipemia, ipoglicemia ed elevato contenuto ematico in acido lattico. L'accumulo di grassi, l'effetto dell'acetaldeide sulle cellule epatiche e altri fattori ancora sconosciuti provocano l'insorgenza dell'epatite alcoolica. La fase successiva è rappresentata dalla cirrosi. La conseguente alterazione della funzione epatica turba la biochimica del sangue, inducendo iperazotemia che può portare al coma e alla morte. Anche la cirrosi modifica la struttura del fegato, ostacolando il flusso del sangue. L'alta pressione esistente nei vasi epatici può indurre la rottura di formazioni varicose e l'accumulo di liquido nella cavità addominale (ascite). Esistono differenze individuali nella reazione all'alcool: in particolare, non tutti i forti bevitori sviluppano epatite e cirrosi.

Composti aromatici policiclici in natura

Questi idrocarburi ad anelli multipli sono stati trovati nel suolo e nei sedimenti in tutto il mondo. Essi sono insolitamente stabili e le loro origini hanno rappresentato un enigma complicato

di Max Blumer

I composti aromatici policiclici sono largamente diffusi in natura; tra essi vi sono pigmenti vegetali e animali di considerevole bellezza e di insolita stabilità chimica. A questa classe di composti appartiene l'alizarina, il colorante rosso delle uniformi militari dell'epoca napoleonica, che mostra tuttora il suo originale splendore nelle collezioni dei musei, dove altri pigmenti vegetali sono sbiaditi da tempo. Animali marini, in particolare echinoidi (ricci di mare) e crinoidi (gigli di mare), devono a tali pigmenti il loro vivace colore. I composti aromatici policiclici sono anche conosciuti per i loro effetti biologici; alcuni di essi possono causare il cancro o comportare mutazioni anche a concentrazioni molto basse.

Allora, che cosa sono i composti aromatici policiclici? Costituiscono una delle principali classi di sostanze in cui la struttura molecolare centrale è tenuta unita da legami stabili carbonio-carbonio. La grande varietà di composti del carbonio - e della vita stessa, che si basa su tali composti - riflette le possibilità quasi illimitate che hanno gli atomi di carbonio di disporsi nello spazio in gruppi, grappoli, catene e anelli. Lunghe catene di atomi di carbonio sono comuni nei prodotti naturali; esse sono state riconosciute già da quando la chimica organica divenne una disciplina a sé. Anche le catene ramificate di atomi di carbonio sono allo stesso modo comuni e hanno una storia scientifica. Il riconoscimento, avvenuto nel 1865, che atomi di carbonio si possono unire per formare anelli chiusi rappresentò uno dei maggiori progressi della chimica. La scoperta fu compiuta da Friedrich Kekulé, che risolse l'enigma della struttura della molecola del benzene dopo aver fatto un sogno in cui serpenti danzanti si mordevano la coda.

Dall'intuizione di Kekulé al riconoscimento che la molecola della naftalina è costituita da due anelli condensati e che le molecole dell'antracene e del fenantrene sono costituite da tre anelli, il passo fu breve. Tali composti ad anelli multipli, o policiclici, sono detti saturi se tutti i legami degli atomi di carbonio, eccettuati naturalmente quelli necessari a costituire il legame carbonio-carbonio, sono uniti ad atomi di idrogeno. Sono detti aromatici se alcuni atomi di carbonio sono legati con doppi legami ad altri atomi di carbonio. Ai tempi di Kekulé si pensava che gli anelli dei composti aromatici avessero legami semplici alternati a legami doppi. Questa concezione tradizionale, ma errata, sopravvive nella rappresentazione grafica abituale dell'anello esagonale del benzene. Secondo le teorie attuali, i legami che uniscono gli atomi di carbonio nella molecola benzenica e in quelle dei composti policiclici hanno tutti lo stesso ruolo e uguale valore. La teoria del legame chimico aiuta a spiegare la stabilità eccezionale dei sistemi ad anello con sei atomi di carbonio.

Gli idrocarburi aromatici policiclici consistono di tre o più anelli benzenici condensati, disposti linearmente, ad angolo o a grappolo. Per definizione, contengono solo atomi di carbonio e di idrogeno. Tuttavia, gli atomi di carbonio dell'anello possono essere facilmente sostituiti da atomi di azoto, zolfo e ossigeno. Ne risultano composti «eteroaromatici» che sono comunemente raggruppati con gli idrocarburi con i quali presentano analogie di proprietà e di comportamento.

L'identificazione degli idrocarburi aromatici policiclici e lo studio del loro corso naturale sono condotti dai geochimici, dai chimici che si occupano dell'ambiente e dai tossicologi. Poiché la quantità di tali idrocarburi è, nella maggior parte dei campioni, piccola, sono necessarie in genere una concentrazione e una separazione preliminari. Nel mio laboratorio alla Woods Hole Oceanographic Institution, usiamo un procedimento a tre fasi che consiste nella filtrazione del gel, nella cromatografia in colonna e nella precipitazione. Questi metodi integragiscono in modi diversi con la miscela di idrocarburi e, combinati, permettono di isolare rapidamente i componenti della miscela. Nella prima fase gli idrocarburi policiclici vengono separati dalle molecole più grosse. Nella seconda, vengono separati dalle sostanze che sono trattenute più o meno fortemente nella colonna cromatografica di allumina. Nella terza fase si forma un precipitato solido con un reagente che si combina in modo specifico con i composti aromatici. Le sostanze estranee vengono escluse dal precipitato e possono essere allontanate mediante lavaggio con un solvente. Questa separazione in tre fasi si effettua con campioni del peso dell'ordine di microgrammi. Per miscele più complesse un'altra colonna cromatografica serve a scindere il concentrato di idrocarburi aromatici in frazioni di molecole aventi uno stesso numero di anelli.

Le frazioni di composti aromatici policiclici ottenute con questa separazione possono essere analizzate con varie tecniche; come regola, nessun metodo può da solo fornire un'analisi completa. Noi abbiamo trovato che si ottiene il maggior numero di informazioni combinando queste tre tecniche: misura dell'assorbimento nell'ultravioletto, separazione per distillazione sotto vuoto spinto e spettrometria di massa. Lo spettro ultravioletto dei campioni di idrocarburi ci dà molte notizie circa la disposizione spaziale degli anelli benzenici in essi contenuti, ma ci fornisce pochi ragguagli sulla presenza e sulla natura dei sostituenti, ossia degli atomi estranei agli anelli stessi. La distillazione scinde successivamente le miscele complesse, semplifica la spettrometria di massa e fornisce informazioni strutturali basate sulle relazioni tra struttura e volatilità. Gli spettri di massa completano gli altri dati; essi misurano la dimensione

Si è trovato che il giglio di mare fossilizzato del genere *Millericrinus* contiene una serie di pigmenti. La fotografia della pagina a fronte mostra la sezione di un peduncolo dell'animale fossile, che visse circa 150 milioni di anni fa sul fondo fangoso del Mar Giurassico, a sud-ovest del quale ora sorge la città di Basilea. Il centro del peduncolo contiene pigmenti cristallizzati chiamati fringeliti. Questo nome deriva da Fringeli, la montagna sulla quale furono trovati i fossili.

totale delle molecole, rivelano la presenza di sostituenti e spesso ne indicano la natura.

Fino a poco tempo fa era sconosciuta la composizione particolareggiata di molte miscele di idrocarburi aromatici policiclici trovati in natura e le origini dei composti erano misteriose. L'analisi chimica può portarci a scoprire i processi di formazione, trasformazione e trasporto dei composti organici in natura? Per rispondere a questa domanda tratterò dapprima alcuni principi che governano la formazione degli idrocarburi aromatici policiclici, dopo di che prenderò in considerazione alcuni casi che rivelano delle correlazioni tra struttura e origini.

Le strutture molecolari dei composti aromatici policiclici si formano ogniqualvolta le sostanze organiche sono esposte ad alte temperature. In questo processo, chiamato pirolisi, viene liberata energia e i prodotti aromatici che si formano sono più stabili dei loro precursori. Per esempio, si ha la pirolisi nel momento in cui un fiammifero viene carbonizzato dalla fiamma. Il carbone che si è formato ha la struttura della grafite: gigantesche molecole piane che consistono di anelli benzenici strettamente uniti fra loro. Tuttavia, per la aromatizzazione e la grafitizzazione della sostanza organica non sono richieste alte temperature e fiamme vive; anche il calore di un ferro da stiro è sufficiente a provocare un'incipiente grafitizzazione e la bruciacchiatura di un tessuto. Dato un tempo sufficiente, l'aromatizzazione procede anche a temperature inferiori. Gli idrocarburi aromatici del petrolio grezzo si sono formati nel corso di milioni di anni in sedimenti che sono stati a temperature comprese tra 100 e 150 gradi centigradi.

La composizione dei prodotti della aromatizzazione termica dipende dalla natura del materiale di partenza e dalle temperature di formazione. Bruciando un ceppo di legno in un camino o un tessuto con un ferro da stiro si ottengono prodotti che sono completamente diversi dal petrolio grezzo, ma che come questo contengono idrocarburi aromatici policiclici. Una caratteristica in particolare dipende dalla temperatura di formazione delle miscele di idrocarburi: l'abbondanza e la distribuzione relativa di idrocarburi aromatici che presentano catene laterali di varie lunghezze, note come gruppi alchilici. A temperature molte elevate, come nella cokizzazione del carbone, i prodotti consistono di una miscela relativamente semplice di idrocarburi non sostituiti, forse a causa della rapida scissione dei legami alchilici che sono meno stabili. A temperature intermedie, come nella combustione lenta del legno, sopravvivono miscele complesse di composti ad anello alchilati. Però anche queste condizioni non sono favorevoli alle lunghe catene alchiliche. Gli idrocarburi non sostituiti superano quantitativamente gli idrocarburi sostituiti e la percentuale degli idrocarburi alchilati diminuisce rapidamente con l'aumentare della lunghezza e del numero delle catene alchiliche.

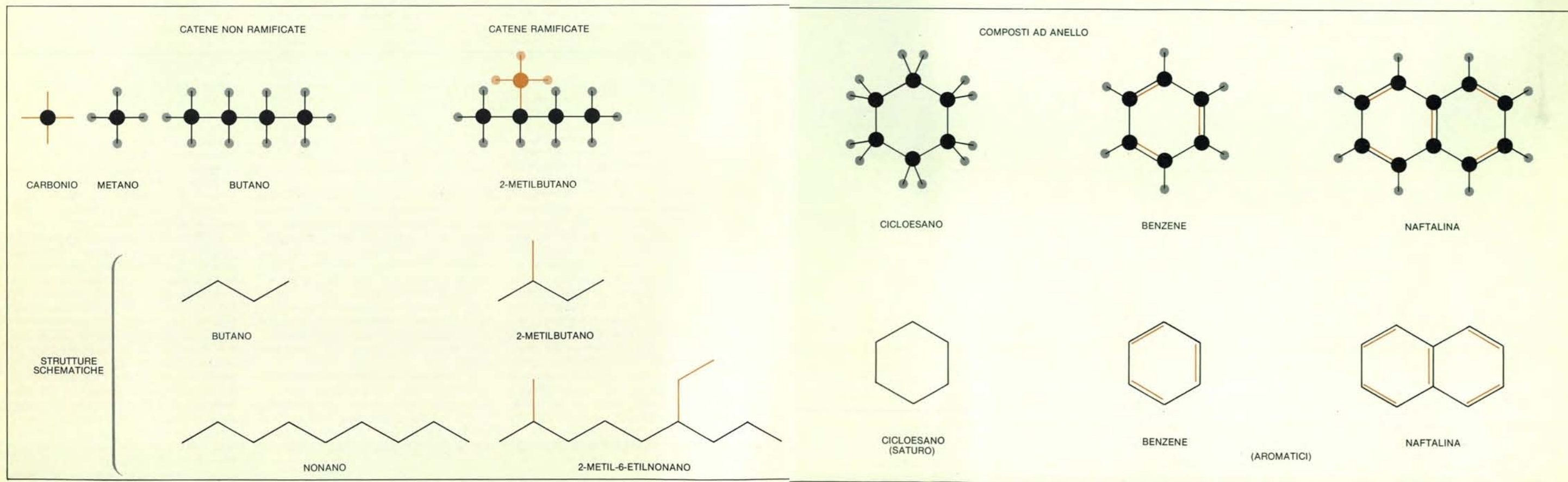
Quando la temperatura di formazione è più bassa si osserva uno schema completamente diverso. Questo è efficacemente illustrato dall'analisi del petrolio grezzo; gli idrocarburi aromatici policiclici alchilati superano di gran lunga gli idrocarburi non sostituiti e il grado medio di alchilazione, come pure il numero massimo degli atomi di carbonio negli anelli aromatici sono molto più elevati di quelli presenti in campioni prodotti per pirolisi ad alta temperatura. Questo fatto riflette le condizioni della formazione del petrolio; il tempo è sufficiente per condurre a termine l'aromatizzazione favorita energeticamente, ma la temperatura non è abbastanza elevata da scindere anche i più deboli legami carbonio-carbonio delle catene alchiliche.

Gli idrocarburi aromatici policiclici si sono formati anche per azione di organismi viventi? Se così fosse, possiamo distinguere questi da quelli originati da processi pirolitici? Molti pigmenti stabili, sia vegetali, sia animali, le cui strutture si fondano su anelli aromatici si sono in realtà formati da organismi viventi. Per definizione, però, tali pigmenti non sono idrocarburi, perché incorporano altri elementi come ossigeno e azoto. Persiste una accesa discussione circa l'ipotesi che gli idrocarburi aromatici policiclici siano anche sintetizzati dagli organismi viventi, direttamente o per trasformazione di precursori che possono somigliare alle strutture esistenti. È particolarmente difficile risolvere sperimentalmente la controversia a causa della ubiquità degli idrocarburi aromatici in natura e della facilità con cui si diffondono e contaminano gli esperimenti.

Al fine di escludere tale contaminazione, Gernot Grimmer del Biochemical Institute for Environmental Carcinogens di Amburgo ha fatto crescere piante in condizioni accuratamente controllate in serre difese da contaminazione esterna. Nonostante l'estesa filtrazione dell'aria, gli idrocarburi aromatici più volatili entrarono nella serra dall'esterno. Tuttavia, sistemi aromatici ad anello, meno volatili, contenenti quattro o più anelli benzenici non sono stati trovati nelle piante, che sembrano impedire la loro biosintesi.

Ci può essere un modo indiretto per risolvere la controversia, o per lo meno possono essere proposti campioni da analizzare come miscele biosintetiche di idrocarburi aromatici policiclici. La biosintesi differisce dalla sintesi pirolitica e geochimica per la sua elevata selettività. La cellula vivente sintetizza solo un numero molto limitato di composti fra tutti quelli teoricamente possibili, e precisamente quelli le cui particolari proprietà sono richieste dall'organismo. Quando numerosi composti appartenenti a una stessa famiglia chimica (per esempio molti idrocarburi a catena lineare) sono formati dalla cellula, spesso sono prodotti in concentrazioni notevolmente diverse. Nella pirolisi e nella geochimica intervengono poche regole di selezione; le miscele risultanti sono straordinariamente complesse e membri adiacenti di famiglie di idrocarburi si trovano in concentrazioni analoghe. Si osserva il caratteristico schema di selezione della biosintesi non solo fra gli idrocarburi a catena lineare e ramificata negli organismi, ma anche fra gli acidi grassi, gli amminoacidi e i carboidrati, ossia fra la maggior parte, se non tutti, dei costituenti della cellula. Questo sembra essere un tale principio fondamentale della biochimica, che ci si aspetta di osservarlo anche in ogni miscela di idrocarburi aromatici policiclici che possa essere prodotta dagli organismi.

In campioni di terreno si trovano composti con un'intensa fluorescenza nella radiazione ultravioletta. Una trentina di anni fa W. Kern, allora assistente presso i laboratori Hoffmann-La Roche di Basilea, si interessò a questi campioni e li studiò nei ritagli di tempo. Nel 1947 scoprì nel terreno di un giardino il crisene, idrocarburo costituito da quattro anelli aromatici e in seguito a questa scoperta divenne il fondatore di una nuova branca della chimica ambientale. Presto furono trovati composti affini. L'isolamento in terreni campestri del benzo[α]pirene, un idrocarburo fortemente cancerogeno, stimolò le analisi in tutto il mondo. Dalla fine degli anni sessanta era convinzione generale che in quasi tutti i terreni, se non in tutti, erano presenti dai 10 ai 15 idrocarburi non sostituiti, ovunque in proporzioni analoghe. Varie ipotesi ne attribuivano la formazione ai batteri del terreno, alla decomposizione di materia vegetale, all'inquinamento da *fallout* atmosferico. Ben presto furono evidenti le limitazioni di questo quadro quando più recenti metodi analitici, dotati di un elevato potere di separazione, furono applicati allo studio di miscele di idrocarburi policiclici nei terreni e nei sedimenti recenti.

Nel terreno e nei sedimenti recenti i più abbondanti sono gli idrocarburi aromatici non sostituiti, che sono accompagnati da numerose serie di membri sostituiti contenenti gruppi metilici, catene alchiliche e composti pentatomici saturi ad anello. Sono anche presenti composti

La varietà dei composti del carbonio è in teoria illimitata. Con le sue quattro valenze, o legami, l'atomo di carbonio forma facilmente catene, ramificate e non, e strutture ad anello in gran varietà. Nei modelli realizzati con palline e bastoncini, illustrati in alto nello schema, le palline grandi rappresentano gli atomi di carbonio e quelle piccole gli atomi di idrogeno. Le strutture ad anello contenenti carbonio, presenti quasi ovunque in natura, sono quelle basate sull'anello benzenico: C_6H_6. Il benzene è descritto come composto insaturo poiché non tutte le valenze libere del carbonio sono legate ad atomi differenti. I chimici usano il temine aromatico per descrivere i composti ad anello insaturi. La naftalina è il più semplice idrocarburo aromatico policiclico, ossia a più anelli. Il cicloesano, C_6H_{12}, è l'analogo saturo del benzene. Nella rappresentazione schematica dell'anello benzenico, illustrata qui, i legami semplici e doppi si alternano nell'anello. In realtà però tutti i legami carbonio-carbonio sono equivalenti come se tra ogni coppia di atomi di carbonio ci fosse un legame e mezzo.

affini agli idrocarburi aromatici, contenenti zolfo. Questi principali elementi strutturali si incontrano in sistemi misti in un numero quasi infinito di permutazioni. La composizione idrocarburica rimane straordinariamente costante su una ampia estensione geografica, dai terreni continentali ai sedimenti che stanno sul fondo marino e dai depositi con caratteristiche ossidanti o fortemente riducenti.

È molto improbabile che la gran diversità di organismi associata a una tale varietà di luoghi possa fornire le stesse serie di idrocarburi nelle stesse proporzioni. La complessità della composizione e l'analogia nelle concentrazioni di composti affini, inoltre, è una prova contro l'origine biochimica.

Allora, può essere valida l'ipotesi di un'origine termica per queste miscele sedimentarie di idrocarburi? Io penso di sì. In tutte le serie predominano gli idrocarburi non sostituiti. Sono presenti derivati alchilici che hanno fino a 10 atomi di carbonio sostituiti, ma la loro concentrazione diminuisce rapidamente con l'aumentare dell'alchilazione (*si veda l'illustrazione in basso a pagina 24*). Questo è esattamente quanto ci si aspetta per la pirolisi a media temperatura. Ci sono

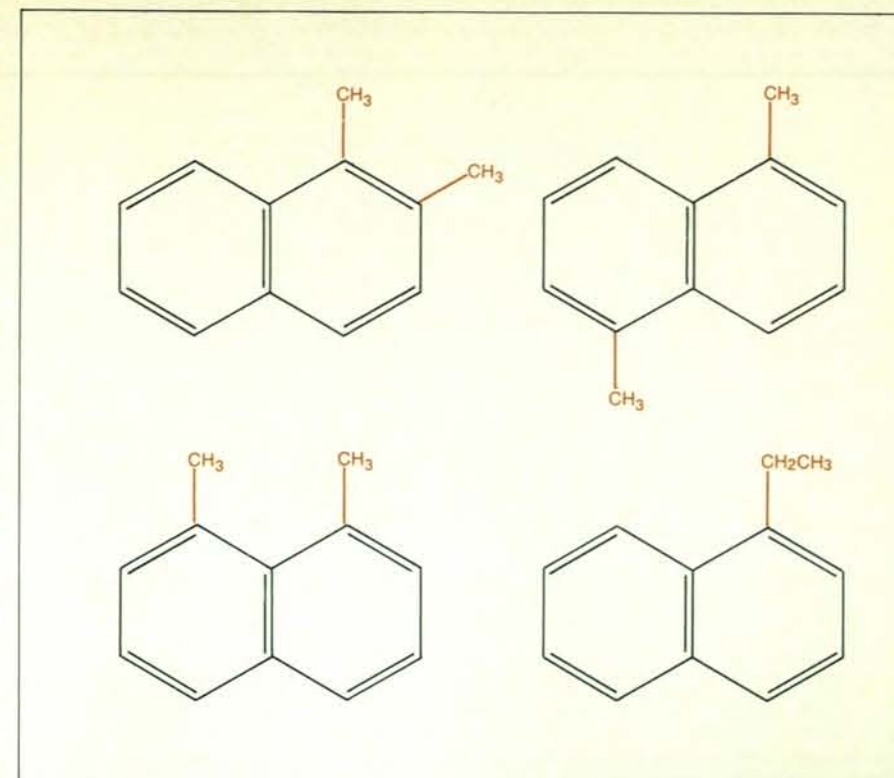
ANTRACENE
TETRACENE
FENANTRENE
CRISENE
PICENE
PIRENE
BENZOPIRENI ISOMERI
PERILENE
CORONENE
MESO-NAFTODIANTRENE

Gli idrocarburi aromatici policiclici consistono di due o più anelli fusi tra loro in vari modi. Gli anelli a sei atomi di carbonio sono i più stabili poiché richiedono la minor distorsione degli angoli naturali di legame degli atomi di carbonio. Gli anelli possono essere disposti linearmente (*prima fila in alto*), ad angolo (*seconda fila*) o a grappolo, come è illustrato dalle restanti configurazioni. I due benzopireni si dicono isomeri perché differiscono nella geometria della molecola. La forma isomerica a destra, benzo[α]pirene, è un agente cancerogeno.

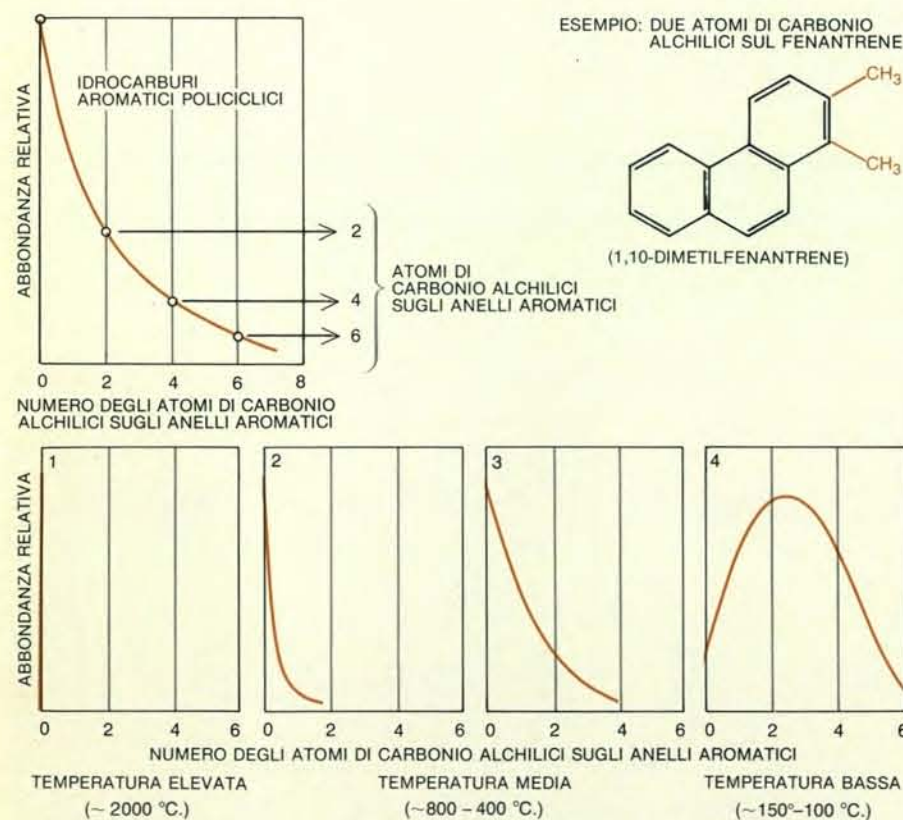
dei prodotti di pirolisi naturale che mostrano uno schema del genere? Il mio collega William W. Youngblood della Florida Technological University ci suggerì di analizzare una sostanza aromatizzante commerciale ottenuta per distillazione del legno di hickory a temperature moderatamente elevate.

Straordinariamente, lo schema di distribuzione alchilica di quel prodotto di pirolisi si accordava con le nostre osservazioni sul suolo e sui sedimenti giovani. Questa prova escludeva la possibilità che numerose fonti termiche avessero contribuito a formare gli idrocarburi sedimentari. Per esempio, gli idrocarburi aromatici policiclici presenti nell'aria inquinata, derivanti dalla combustione incompleta nelle fornaci e nei motori a temperature più elevate, contengono meno derivati alchilici di quelli che si trovano nei campioni di sedimento. Il petrolio grezzo, all'altra estremità dello spettro delle temperature di formazione, è molto più alchilato. Se gli idrocarburi policiclici si fossero formati per decomposizione di materiale vegetale, dovrebbero essere ancora più alchilati. In ogni caso si deve considerare improbabile che la decomposizione di materiale vegetale concorra a formare la frazione di idrocarburi aromatici policiclici trovata in tutto il mondo nei sedimenti per il motivo che ho già citato: l'enorme diversità degli organismi coinvolti. Pertanto, l'evidenza suggerisce che la frazione di idrocarburi aromatici policiclici viene prodotta mediante processi pirolitici naturali. Ma come è possibile che la composizione della frazione sia così costante in una serie di campioni tanto ampia?

Io ho una teoria interessante che spiegherebbe la relativa uniformità della frazione di idrocarburi aromatici policiclici trovati nel suolo e nei sedimenti marini recenti. Grandi quantità di prodotti di pirolisi si formano in seguito a incendi che si sviluppano nelle foreste e nelle praterie e poi vengono dispersi dai venti. La caligine che sovrasta l'Atlantico settentrionale è attribuita in parte a tali incendi. In realtà, Dwight M. Smith, John J. Griffin e Edward D. Goldberg della Scripps Institution of Oceanography hanno trovato particelle di carbone con una riconoscibile struttura legnosa nei sedimenti marini profondi. Altri ricercatori hanno riferito di aver rinvenuto carbonio elementare nei depositi marini di manganese dove noi avevamo già scoperto idrocarburi aromatici.

Oltre a tali particelle di carbone, gli incendi delle foreste e delle praterie producono in grande abbondanza idrocarburi aromatici. Come la massa di gas, raffreddandosi, sale dagli incendi, le particelle attive di carbone che si sono appena formate possono catturare gli idrocarburi aromatici e proteggerli dall'ossidazione indotta dalla luce durante il loro movimento attraverso l'atmosfera. La miscelazione che avviene durante il trasporto potrebbe spiegare sia l'uniformità di composizione su una vasta area, sia il motivo per cui gli organismi e le condizioni chimiche del luogo dove si deposi-

Gli isomeri degli alchilnaftaleni contengono per definizione lo stesso numero di atomi di carbonio e di idrogeno. Esistono solo quattro isomeri dei numerosi possibili. Il termine alchile si riferisce alle catene laterali. Benché gli isomeri siano identici chimicamente le loro differenze strutturali danno luogo a proprietà diverse. Nonostante ciò, è difficile isolare gli isomeri in forma pura.

Il primitivo pigmento rosso fu estratto dall'autore dal peduncolo fossilizzato del giglio di mare del genere *Millericrinus* (*si veda l'illustrazione a pagina 18*). Il pigmento appartiene alla famiglia delle fringeliti che differiscono fra loro per il numero e la posizione dei gruppi ossidrile (OH).

L'aggiunta di atomi diversi agli idrocarburi aromatici policiclici aumenta notevolmente il numero delle permutazioni strutturali possibili. L'alizarina, il brillante pigmento rosso dell'epoca napoleonica, contiene nella sua molecola atomi di ossigeno e gruppi ossidrilici come sostituenti periferici. Presenta una certa somiglianza con la struttura della fringelite illustrata nella pagina precedente in basso.

Il numero degli atomi di carbonio alchilici presenti come catene laterali negli idrocarburi aromatici policiclici è strettamente legato alla temperatura alla quale si formarono i composti. In una miscela tipica (*curva in alto a sinistra*) gli anelli che non hanno catene laterali sono i più comuni; in generale, più sono gli atomi di carbonio nelle catene laterali, minore è l'abbondanza. Un esempio di composto policiclico con due atomi di carbonio alchilici è l'1,10-dimetilfenantrene. Quando i composti aromatici si formano a temperatura elevata (*1*), le catene laterali sono virtualmente assenti. A temperature inferiori (*2,3*), l'abbondanza di catene laterali aumenta con la diminuzione della temperatura. Tuttavia, alle basse temperature che si ritiene siano associate con la formazione della maggior parte del petrolio, i composti policiclici con due o tre atomi di carbonio alchilici superano numericamente le altre configurazioni di idrocarburi policiclici.

tano non si riflettono sulla composizione chimica delle miscele di idrocarburi. Da ultimo, l'ampia estensione delle temperature di combustione negli incendi naturali spiegherebbe la varietà dei derivati alchilati osservata.

Alcuni rari minerali organici, dotati di una bella fluorescenza, sono stati rinvenuti in California e nell'Europa orientale; si trovano insieme a minerali di mercurio e a volte con acque minerali e sugli sfiati di gas infiammabili. Lo studio di questi minerali risale al chimico francese Jean-Baptiste-André Dumas, che vi lavorò nella prima metà del XIX secolo. Intorno al 1880, chimici tedeschi isolarono due idrocarburi aromatici costituiti da quattro anelli in campioni prelevati in miniere di mercurio a Idrija in Iugoslavia. La idrialite proveniente da quella località, la curtisite proveniente da Skaggs Springs, in California, e minerali affini trovati in altre località sono stati da tempo oggetto di numerose ricerche. I metodi analitici semplici usati fino in tempi recenti suggerirono una composizione semplice con pochi componenti, forse solo uno. Come con gli idrocarburi sedimentari, i nostri metodi più nuovi hanno rivelato una complessità di composizione del tutto imprevista. Benché la idrialite e la curtisite siano nettamente differenti dal punto di vista chimico, le serie di idrocarburi in esse presenti si sovrappongono per un certo grado. Esse contengono almeno diverse centinaia di idrocarburi aromatici policiclici, insieme a composti analoghi contenenti zolfo e azoto e derivati alchilici e cicloalchilici in molte combinazioni di sostituzione.

L'analisi chimica può dirci qualcosa circa l'origine di questi minerali? La distribuzione delle catene alchiliche degli idrocarburi aromatici policiclici nella idrialite e nella curtisite assomiglia alla distribuzione delle stesse nel suolo e nei sedimenti recenti. Pertanto gli idrocarburi non sostituiti sono i più abbondanti. Tuttavia le serie alchiliche non si estendono tanto, e la diminuzione della concentrazione da un membro di una serie al successivo è maggiore che non nel distillato del legno di noce. Ciò suggerisce l'ipotesi di un'origine pirolitica a temperature più elevate di quelle degli incendi nelle foreste e nelle praterie, ma inferiori a quelle dei forni e dei motori.

L'insieme degli idrocarburi policiclici presenti nella idrialite e nella curtisite differisce stranamente da quello trattato finora per il modo in cui sono disposti gli anelli benzenici. In base alla disposizione degli anelli, si hanno tre ampie categorie: lineare (con tutti gli anelli su una linea), angolare (con gli anelli a scala) e a grappolo (con almeno un anello circondato su tre lati), che presentano diversa stabilità. La disposizione lineare nell'antracene e nel tetracene è la meno stabile; il tetracene e gli idrocarburi più complessi, a esso equivalenti, possono essere preparati in laboratorio, ma non sopravvivono in natura. Il pirene, il benzopirene e il perilene, che sono costituiti da grappoli di anelli benzenici, sono più stabili e si trovano comunemente nei prodotti di pirolisi. La configurazione più stabile è quella in cui gli anelli benzenici sono disposti ad angolo, come nel fenantrene, nel crisene e nel picene. Tali serie stabili abbondano nella idrialite e nella curtisite, che contengono pochi idrocarburi del tipo a grappolo che sono comuni nel suolo, nei sedimenti, nel catrame di legno, nel fumo del tabacco, nel petrolio e nei gas di scarico delle automobili.

Evidentemente, le serie del tipo a grappolo si formano nella pirolisi indipendentemente dalla temperatura. Permangono se i prodotti di reazione vengono rapidamente raffreddati a temperature in cui si arresta la decomposizione dei prodotti meno stabili, come nel caso del catrame di legno; si trovano anche nel petrolio poiché quest'ultimo non ha mai raggiunto temperature favorevoli alla eliminazione o alla ristrutturazione dei sistemi ad anello meno stabili.

Ciò suggerisce che la idrialite e la curtisite perdono le configurazioni a grappolo instabili, poiché queste sono state eliminate per effetto della prolungata esposizione alle alte temperature.

Con tali conoscenze del comportamen-

Un alone di grafite si forma attorno a cristalli di chiastolite (silicato di alluminio) in crescita, in alcuni minerali associati a sedimenti sepolti in profondità. In questi sedimenti la materia organica è trasformata in composti organici che hanno volatilità diversa. La frazione più refrattaria è la grafite, che rimane in esemplari come quello raffigurato qui e rinvenuto da R. Sawdo della Woods Hole Oceanographic Institution.

to chimico e con le nozioni geologiche generali si possono interpretare la formazione e l'evoluzione di questi minerali idrocarburici in questo modo. I sedimenti contenenti composti organici vengono trasportati dai movimenti della crosta terrestre verso regioni più profonde di quelle in cui il petrolio si è tipicamente formato. In tali regioni più profonde la temperatura alla quale avviene la pirolisi raggiunge i 400-500 gradi centigradi. Il materiale organico originale viene distrutto: i suoi costituenti subiscono una ristrutturazione e diventano termodinamicamente stabili. Uno dei prodotti è la grafite, che rimane alla profondità originaria. I gas idrocarburici stabili (in particolare il metano) e gli idrocarburi aromatici policiclici (insieme con i composti analoghi contenenti zolfo e azoto) rimangono press'a poco alla profondità originaria per un tempo abbastanza lungo perché siano distrutti gli idrocarburi meno stabili. Alla fine il resto degli idrocarburi si muove verso la superficie con i gas, le acque minerali e i minerali di mercurio che sono riconosciuti per la loro mobilità geochimica. Durante la migrazione, gli idrocarburi si separano per cristallizzazione frazionata. Le frazioni a peso molecolare e punto di fusione più elevati formano la idrialite e la pendletonite; le frazioni aventi peso molecolare e punto di fusione più bassi danno luogo alla curtisite.

L'analisi chimica ci ha nuovamente fornito la chiave della formazione di un insieme naturale di idrocarburi; essa suggerisce grossolanamente la modalità di formazione e fornisce una prova circa la temperatura di formazione e la durata dell'esposizione al calore. Col proseguire delle nostre ricerche, noi trovammo una spiegazione per la formazione di idrocarburi in materiali così diversi come terreni, sedimenti recenti, catrame di legno, fumo di tabacco, gas di scarico dei motori e minerali quali la idrialite e la curtisite. Studiando gli idrocarburi aromatici policiclici presenti nel petrolio, trovammo anche correlazioni fra la struttura chimica e i processi di formazione degli idrocarburi.

Il petrolio può rappresentare la miscela organica più complessa presente sulla Terra. Esso si è formato dai residui della vita primordiale in sedimenti sepolti, attraverso reazioni chimiche che richiedono milioni di anni per completarsi. La trasformazione assomiglia alla pirolisi, ma le reazioni sono straordinariamente lente a causa delle moderate temperature coinvolte: forse inferiori ai 150 gradi centigradi. Si formano così e si conservano composti che non si sarebbero osservati nei prodotti della pirolisi a elevate temperature.

L'analisi chimica della frazione di idrocarburi aromatici policiclici presente nel petrolio è una specie di sfida. Un'adeguata scissione analitica richiede la combinazione di molte tecniche diverse; inoltre raramente vengono isolati composti semplici puri. Harold J. Coleman e i suoi collaboratori al Bartlesville Energy Research Center di Oklahoma hanno condotto ampie analisi sulle frazioni aromatiche policicliche del petrolio grezzo, in particolare su quelle provenienti dalla regione di Prudhoe Bay in Alaska. Gli idrocarburi aromatici policiclici sono abbondanti in quanto ammontano a circa un sesto della frazione di petrolio che distilla tra 370 e 535 gradi centigradi. Una proporzione simile è presente in alcune frazioni che hanno un punto di ebollizione più basso, nei distillati superiori e nel residuo di distillazione. Tre caratteristiche di composizione sono qui particolarmente adeguate: l'estrema complessità e l'uniformità nella concentrazione di composti adiacenti entro la numerosa serie di idrocarburi, la frequente incidenza di composti con anelli pentatomici saturi e la presenza di composti in cui i gruppi sostituenti mostrano una considerevole tensione sterica, ossia presentano angoli di legame distorti oltre il loro limite normale.

CICLIZZAZIONE

AROMATIZZAZIONE

FORMAZIONE DI ANELLI PENTATOMICI

La ciclizzazione e l'aromatizzazione di catene idrocarburiche sature danno luogo a miscele complesse di composti aromatici policiclici trovati nel petrolio. Il materiale d'origine del petrolio, derivato in gran parte dalle piante, è ricco di lunghe catene di atomi di carbonio. A elevate temperature nei sedimenti sepolti, alcune catene perdono pochi atomi di idrogeno e si trasformano in anelli esatomici (*in alto*). Per successiva perdita di idrogeno, gli anelli saturi si trasformano in aromatici (*in centro*). La sostituzione con catene, seguita da ciclizzazione e aromatizzazione, forma gradualmente composti con molti anelli. Gli anelli pentatomici, che pure si formano facilmente (*in basso*), non possono essere trasformati in anelli aromatici. Se attorno a un nucleo aromatico se ne accumulano diversi, si blocca la crescita successiva e quindi la formazione di un sistema aromatico più ampio. Nel petrolio abbondano idrocarburi aromatici circondati da numerosi anelli pentatomici. La loro produzione è favorita da un tempo di reazione di milioni di anni e dall'assenza di temperature troppo elevate.

La complessità e l'uniformità nella concentrazione di composti adiacenti caratterizzano l'origine di questa frazione di petrolio grezzo; non è quindi un modello biochimico, ma il risultato di un ampio rimescolamento geochimico di composti organici mediante reazioni non selettive. Nella formazione del petrolio, alcune strutture lineari del carbonio sono trasformate in anelli saturi a cinque e a sei atomi. Tali reazioni sono possibili poiché rendono termodinamicamente stabile il materiale di origine. Gli anelli saturi a sei atomi così creati vengono subito trasformati per aromatizzazione - perdita di idrogeno - in sistemi aromatici ancora più stabili. Tali sistemi possono aumentare mediante l'accrescimento e l'aromatizzazione di nuovi anelli esatomici saturi.

Gli anelli a cinque atomi, d'altra parte, non si convertono subito in strutture aromatiche, così che con il passare del tempo due, tre o più anelli possono crescere attorno a un nucleo aromatico, bloccando così il suo successivo sviluppo in un grosso sistema aromatico. Gli anelli pentatomici attorno a un nucleo aromatico sono numericamente più abbondanti nel petrolio che nelle serie idrocarburiche formatesi a elevate temperature. Ciò riflette il fatto che il petrolio ha avuto a disposizione per la sua formazione molto più tempo; l'ordinamento degli anelli può continuare, sebbene a bassa intensità, per milioni di anni. Nella pirolisi a temperature più elevate gli anelli pentatomici contenuti nelle strutture aromatiche sono meno abbondanti sia perché il tempo di reazione è breve, poiché il materiale di partenza è rapidamente trasformato in composti a sei atomi e in sistemi aromatici, sia perché la temperatura è abbastanza elevata da rompere i legami carbonio-carbonio presenti negli anelli saturi.

Molecole sotto forte tensione in cui un singolo gruppo CH_2 forma un ponte tra due atomi di carbonio aromatici, come nel 4,5-metilenfenantrene (*si veda l'illustrazione in questa pagina*), si formano con abbondanza relativa diversa in vari pirolisati. Essi sono abbondantissimi nel catrame di legno, meno abbondanti nel terreno e nei sedimenti e ancora meno abbondanti nel petrolio grezzo. Nella idrialite e nella curtisite tali composti sono pressoché assenti. L'angolo di legame distorto al ponte CH_2 implica che tali molecole abbiano un contenuto energetico più elevato delle molecole con tensione minore e pertanto siano meno stabili. Noi attribuiamo la formazione di ponti CH_2 sotto tensione nel corso della pirolisi a reazioni energetiche che resistono al raffreddamento rapido (come nel catrame e nel fumo) o alle basse temperature (come nella formazione del petrolio) che abbassano la velocità alla quale può avvenire una ristrutturazione. I ponti CH_2 non sono presenti nella idrialite e nei minerali simili perché furono esposti a temperature elevate per lunghi periodi dopo la pirolisi iniziale.

Anelli pentatomici sotto tensione si formano per pirolisi (esposizione a temperature elevate) da catene alchiliche laterali. Se la struttura madre fornisce tre dei cinque lati dell'anello finale è necessario un solo atomo di carbonio alchilico (*a sinistra*). Negli altri casi (*a destra*) possono essere necessarie due catene alchiliche. Poiché i legami carbonio-carbonio sono sotto una tensione maggiore negli anelli pentatomici che non in quelli esatomici, i primi hanno un contenuto energetico maggiore e pertanto una minore stabilità. Gli anelli pentatomici sopravvivono senza grande difficoltà se la miscela pirolitica viene prontamente raffreddata, come avviene nel caso del fumo prodotto da un incendio, o se la pirolisi continua per lungo tempo a bassa temperatura.

Questi esempi serviranno a dimostrare l'abbondanza di informazioni che sono racchiuse nella struttura di composti organici presenti in natura. I parametri che ho trattato riflettono la caratteristica dei materiali di partenza, i processi di formazione, le temperature alle quali i composti si formano e il tempo di reazione. Le relazioni tra questi parametri suggeriscono che questi composti potrebbero essere usati come «termometri» e «orologi» geologici. Tuttavia la loro «taratura» presenta un problema difficile. Ci deve essere un'interdipendenza complessa tra i parametri, poiché il tempo e la temperatura influenzano entrambi la sopravvivenza di strutture instabili. Tuttavia, anche allo stadio attuale delle nostre conoscenze, l'analisi chimica particolareggiata ci ha aiutato a riconoscere una ampia gamma di processi ambientali e ci ha guidato verso una visione unificata di materiali che sono remoti per quanto riguarda l'origine e i processi di formazione.

Non tutte le miscele naturali di idrocarburi aromatici policiclici hanno un'unica origine predominante. Spesso si presenta una situazione più complessa in cui più di una fonte di idrocarburi contribuisce alla formazione del campione. Per esempio, un sedimento marino recente può contenere combustibili fossili derivati da petrolio versato in mare in aggiunta agli idrocarburi che noi attribuiamo alla caduta di particelle di fuliggine provenienti dagli incendi di foreste e praterie. I differenti schemi di composizione dovrebbero rendere possibile il ri-

conoscimento di queste varie origini.

Con il mio collaboratore Jeremy Sass ho studiato un caso analogo. Uno spargimento di petrolio nella Buzzards Bay poco lontana dal Massachusetts ha contaminato il sedimento profondo vicino alla costa con un olio pesante che ha un punto di ebollizione relativamente basso. L'estrazione e l'isolamento della frazione contenente idrocarburi aromatici policiclici a punto di ebollizione più elevato e le analisi successive rivelano lo schema di base normale che osserviamo non solo in ogni parte della baia, ma anche nel terreno a essa adiacente. Gli aromatici non sostituiti predominano e lo schema di distribuzione dei derivati alchilici è la sola caratteristica del fumo di legna, cioè degli idrocarburi che noi attribuiamo alla ricaduta di fuliggine. Dentro la gamma dei punti di ebollizione dei composti presenti nel combustibile versato, tuttavia, i derivati alchilici sono più abbondanti degli idrocarburi non sostituiti. Questa è una caratteristica del petrolio ed è straordinario che possa essere ancora osservata adesso, quasi sei anni dopo.

Un esempio particolarmente interessante di una doppia origine deriva da una nostra più recente ricerca di idrocarburi in un crinoide fossile proveniente dalla catena del Giura in Svizzera. Questo animale marino fossilizzato ha una intensa colorazione prodotta da pigmenti che sono dispersi nella matrice rocciosa del fossile. Il pigmento, presente sotto forma di minuscoli cristalli, è un minerale organico vero e proprio; l'abbiamo chiamato fringelite dal nome della montagna, Fringeli, dove fu scoperto. Poiché composti analoghi si trovano tuttora in esemplari viventi affini a questa specie animale, noi presumiamo che il pigmento, o almeno i suoi stretti precursori, facesse parte dell'animale che visse nel Mare Giurassico 150 milioni di anni fa.

La fringelite e le sostanze a essa affini sono composti aromatici. Quando tentammo di scindere i pigmenti fossili mediante cromatografia notammo nella prima frazione che usciva dalla colonna dei pigmenti dei materiali quasi incolori, ma intensamente fluorescenti. Studi successivi rivelarono una serie continua di composti policiclici che contenevano da tre a sette anelli aromatici. I membri inferiori della serie non sono particolarmente abbondanti e assomigliano molto alle miscele di composti aromatici che si trovano in molti altri esemplari geologici. Possono essere ciò che è rimasto di composti che furono sintetizzati in ere geologiche passate o, ciò che è più probabile, possono essere i prodotti di una trasformazione geologica che ha avuto origine nei sedimenti profondi non sufficientemente ricchi di materiale organico per formare petrolio. A pesi molecolari più elevati trovammo idrocarburi insoliti, che non

CARATTERISTICHE STRUTTURALI	PREFERENZA STRUTTURALE	CONDIZIONE DI FORMAZIONE
GRADO DI ALCHILAZIONE	BASSO GRADO DI ALCHILAZIONE	ALTA TEMPERATURA DI FORMAZIONE ALTO CONTENUTO DI CARBONIO NEL MATERIALE DI ORIGINE
DISPOSIZIONE DEGLI ANELLI (LINEARE, AD ANGOLO, A GRAPPOLO)	DISPOSIZIONE AD ANGOLO DEGLI ANELLI	TEMPO LUNGO DI EQUILIBRAMENTO ELEVATA TEMPERATURA DI EQUILIBRAMENTO
TENSIONE	ALTO GRADO DI TENSIONE NELL'ANELLO	TEMPO BREVE DI EQUILIBRAMENTO BASSA TEMPERATURA DI FORMAZIONE
NUMERO DI ANELLI PENTATOMICI	ANELLI PENTATOMICI MULTIPLI	TEMPO LUNGO DI REAZIONE BASSA TEMPERATURA DI FORMAZIONE

La configurazione delle strutture idrocarburiche trovate nelle miscele geochimiche fornisce preziose informazioni sulle condizioni in cui le strutture si formarono. Il tipo e l'ordinamento spaziale delle catene alchiliche e degli anelli saturi e aromatici può variare in un ampio raggio. La presenza di strutture privilegiate in miscele naturali di idrocarburi aromatici policiclici riflette la composizione del materiale d'origine, le temperature di formazione e di trasformazione e la durata dei processi chimici coinvolti nella formazione delle varie molecole.

erano mai stati isolati da sedimenti e che presentavano proprio quelle caratteristiche che noi avevamo previsto per sostanze di origine biochimica, piuttosto che geochimica. Il numero di tali composti è limitato, le strutture chimiche sono insolite e le concentrazioni sono molto più elevate del normale (*si veda l'illustrazione qui a destra*).

Un'analisi strutturale più dettagliata sostiene questa interpretazione. Gli idrocarburi insoliti sono i sistemi aromatici ad anello sui quali si basano le fringeliti. Non siamo sicuri se quegli idrocarburi fossero già presenti negli animali vivi e pertanto rappresentino puramente prodotti biochimici, oppure se siano prodotti di trasformazione geochimica di precursori biochimici strettamente affini alle fringeliti. Entrambe le teorie sosterrebbero il valore profetico delle relazioni che ho suggerito tra struttura chimica e origine biologica.

La nostra visione generale degli idrocarburi aromatici presenti in natura è ora cambiata in modo sostanziale. Il quadro di circa una dozzina di composti semplici distribuiti sulla Terra in proporzioni simili ci ha dato modo di dedurre che la miscela è straordinariamente complessa e mutevole e che le sue origini sono diverse. Finora la miscela non è stata analizzata completamente; in realtà, allo stato attuale delle conoscenze in campo analitico, non può essere scomposta in tutti i suoi componenti. Finora le nostre analisi, benché incomplete, rivelano l'intervento di numerosi processi chimici fondamentali e noi stiamo imparando a leggere nelle strutture idrocarburiche un codice che può essere decifrato in termini di origini e di processi di formazione, trasformazione e movimento di massa. Abbiamo esaminato un'ampia gamma di campioni e la prova che gli idrocarburi aromatici policiclici naturali abbiano una origine termica è schiacciante.

Idrocarburi aromatici policiclici derivati da pirolisi sono stati presenti sulla Terra per lungo tempo. L'uomo è sempre stato a contatto con prodotti di combustione e gli incendi naturali e le reazioni nei sedimenti formarono gli idrocarburi aromatici policiclici molto prima della comparsa dell'uomo. Tuttavia, adesso sappiamo che gli idrocarburi presenti nel fumo, nel *fallout* atmosferico, nei sedimenti e nei combustibili fossili, contengono, oltre a composti già riconosciuti che danno luogo al cancro e a mutazioni, nuovi composti cancerogeni e mutageni. Da ciò deriva una nuova domanda: tali materiali hanno contribuito in modo significativo al ruolo di mutazione nell'evoluzione della specie? Essi potrebbero essere classificati con gli altri mutageni naturali come la radiazione ultravioletta e la radiazione nucleare.

Gli idrocarburi aromatici policiclici ora entrano a far parte dell'ambiente in quantità maggiori che non nelle passate ere geologiche.

Tra le origini predominanti figurano la combustione incompleta del legno, del carbone e del petrolio e lo scarico di petrolio grezzo o raffinato. Sono state condotte prove per valutare gli effetti di questo nuovo afflusso di idrocarburi aromatici nell'ambiente, tuttavia queste prove sono precedenti al recente riconoscimento che la miscela ambientale di idrocarburi aromatici è straordinariamente complessa, che ha molte origini, che è difficile da analizzare e che di molti componenti non è mai stata saggiata l'attività biologica. Sembra importante riprendere in esame il problema degli effetti sull'ambiente del nuovo afflusso di sostanze aromatiche alla luce delle nostre cognizioni attuali e di ciò che ancora non conosciamo.

Nel giglio di mare fossile *Millericrinus* si trovano due gruppi di idrocarburi. Il gruppo meno abbondante consiste di strutture abbastanza piccole da tre a sei anelli benzenici condensati trovati in abbondanza anche in sedimenti antichi. Il gruppo più numeroso consiste di strutture policicliche molto più grandi (*in colore*) che non sono mai state trovate in altri depositi. Le loro strutture sono così simili a quelle dei pigmenti detti fringeliti che sembra possibile un'origine comune. È questo uno dei pochi casi in cui una parte di una miscela naturale di idrocarburi aromatici policiclici non ha origine da pirolisi, ma può essere fatta risalire a un organismo vivente.

Dal primo studio di Dumas sulla idrialite e dalla scoperta, dovuta a Kern, del crisene nel terreno, la nostra conoscenza degli idrocarburi aromatici policiclici si è sviluppata di pari passo con l'analisi: lentamente durante i periodi in cui l'analisi subì una stasi, rapidamente con la convergenza di molti metodi analitici moderni. Io noto uno sviluppo analogo in altri campi della chimica organica ambientale. Ciascuna delle nostre tecniche migliori è insufficiente da sola per ottenere la scissione completa di composti chimici. La domanda: «Quanto complessa è la natura?» è importante e ha molte implicazioni. Come geochimico, vedo un'opportunità quasi illimitata a interpretare un mondo complesso in base ai processi che lo hanno formato. Come biologo ambientale, sono frustrato dalla difficoltà, se non dalla impossibilità, di prevedere l'effetto dei composti organici in natura senza una più completa conoscenza della loro struttura.

La meteorologia di Giove

Le caratteristiche visibili del pianeta gigante riflettono la circolazione della sua atmosfera. Un modello che riproduca tali caratteristiche dovrebbe essere valido anche per atmosfere di altri pianeti, compresa la Terra

di Andrew P. Ingersoll

Tutto quello che si riesce a vedere dalle fotografie del pianeta Giove sono nuvole: le fasce scure, le zone più chiare e la grande macchia rossa. La superficie solida, se davvero esiste, giace molte migliaia di chilometri al di sotto della superficie visibile. Eppure la maggior parte delle caratteristiche atmosferiche di Giove persiste per moltissimo tempo e possiede una struttura organizzata sconosciuta all'atmosfera terrestre. Queste differenze, assieme al fatto che è facile osservare le caratteristiche dell'atmosfera di Giove, fanno del pianeta gigante un laboratorio in cui i meteorologi terrestri possono verificare le teorie sulla dinamica dell'atmosfera in maniere impossibili sulla Terra.

Il diametro di Giove è 11 volte più grande di quello della Terra: la gravità alla sua superficie è 2,4 volte più grande, la sua velocità di rotazione è 2,4 volte più grande (una rotazione completa in 10 ore). Su Giove le variazioni stagionali sono quasi inesistenti, perché il suo asse di rotazione è quasi parallelo a quello della sua orbita attorno al Sole.

L'atmosfera e l'interno di Giove sono composti principalmente d'idrogeno, mentre altri elementi, quali elio, carbonio, ossigeno e azoto vi sono mescolati nelle stesse proporzioni in cui sono presenti nel Sole. Poiché questa miscela non solidifica nelle condizioni di pressione e temperatura che sono state calcolate per Giove, probabilmente il pianeta è completamente gassoso o liquido. A tutte le profondità la miscela viene lentamente mescolata da correnti convettive che portano calore dall'interno alla superficie. Calcoli teorici indicano che molto probabilmente il calore interno di Giove è dovuto ancora all'energia gravitazionale della sua contrazione iniziale, di quando, cioè, Giove si formò per condensazione della stessa nebulosa che diede origine al Sole e agli altri pianeti.

Dal punto di vista della meteorologia le differenze principali fra Giove e la Terra consistono nel fatto che Giove ha una notevole fonte di energia interna e manca probabilmente di una superficie solida. Le altre differenze - di raggio, gravità, velocità di rotazione e così via - sono essenzialmente quantitative. Perfino la composizione chimica dell'atmosfera di Giove è simile alla composizione di quella terrestre, per quello che riguarda i suoi effetti sulla meteorologia. Le atmosfere dei due pianeti consistono essenzialmente di gas non condensabili: idrogeno ed elio su Giove, azoto e ossigeno sulla Terra; in piccole quantità vi sono mescolati vapore acqueo e altri gas che possono condensare e formare nubi. Dal punto di vista della variazione di temperatura che si avrebbe sui due pianeti se tutti i vapori condensabili passassero allo stato liquido o solido, cedendo così tutto il loro calore latente, l'atmosfera di Giove sarebbe influenzata dalla condensazione un pochino meno di quella della Terra. In ogni caso le nubi, la condensazione e la precipitazione sono importanti nella dinamica di ambedue queste atmosfere.

Il quadro che abbiamo della composizione e della struttura verticale dell'atmosfera gioviana si basa in parte su osservazioni e in parte sulla teoria. Studi spettroscopici compiuti sulla Terra hanno stabilito che l'atmosfera è composta principalmente di idrogeno molecolare (H_2) e di quantità minori di metano (CH_4), di ammoniaca (NH_3) e di un elenco di altri gas che si va allungando. In questi studi si paragonano le caratteristiche di assorbimento dello spettro infrarosso di Giove con le caratteristiche dello spettro dei gas in laboratorio. Poiché ogni gas ha delle lunghezze d'onda caratteristiche alle quali assorbe radiazioni, gli spettri di assorbimento dell'atmosfera gioviana rendono possibile identificare con sicurezza la maggior parte dei gas presenti in concentrazioni anche piccolissime. Fa eccezione l'elio, che assorbe radiazioni solo nella parte ultravioletta dello spettro, a lunghezze d'onda che non si possono osservare attraverso l'atmosfera terrestre. Tuttavia recentemente l'elio è stato determinato per mezzo di uno spettrometro ultravioletto posto sulla navicella spaziale *Pioneer 10*, e si è osservato anche l'effetto dell'elio sullo spettro infrarosso di altri gas presenti.

Dalle intensità relative degli spettri di assorbimento dell'idrogeno, del metano e dell'ammoniaca si possono ricavare le concentrazioni relative degli stessi gas. Su Giove il rapporto fra il numero di atomi di carbonio e gli atomi di idrogeno è circa 1/3000 e il rapporto atomi di azoto su atomi di idrogeno è 1/10 000. Queste abbondanze sono prossime a quelle degli stessi elementi nel Sole. Entro certi limiti il rapporto elio-idrogeno su Giove è simile al rapporto elio-idrogeno ricavato per il Sole (1/15). Sono proprio i rapporti di concentrazione, insieme alla bassa densità di Giove nel suo insieme, che fanno pensare che la sua composizione sia molto simile a quella del Sole.

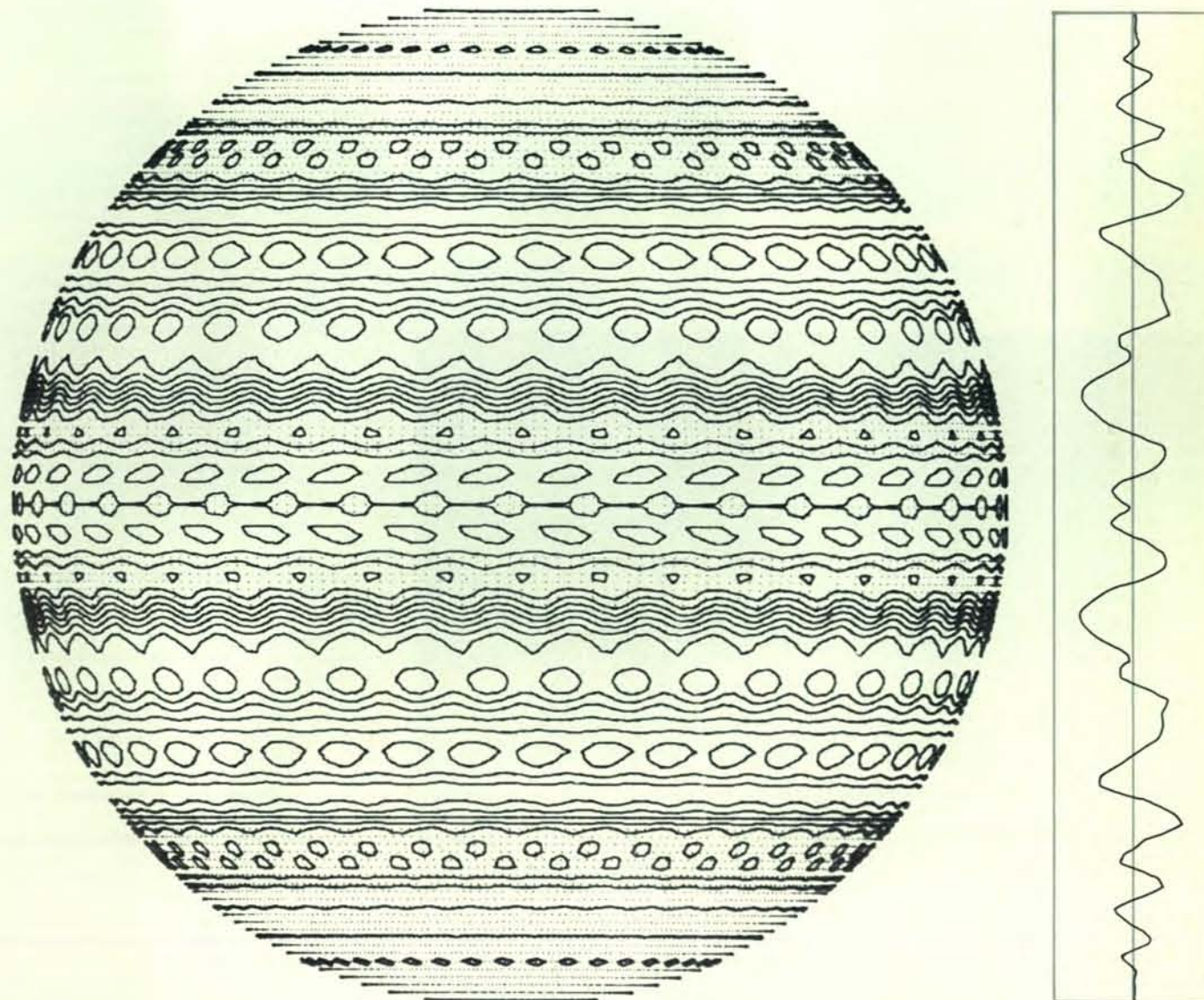
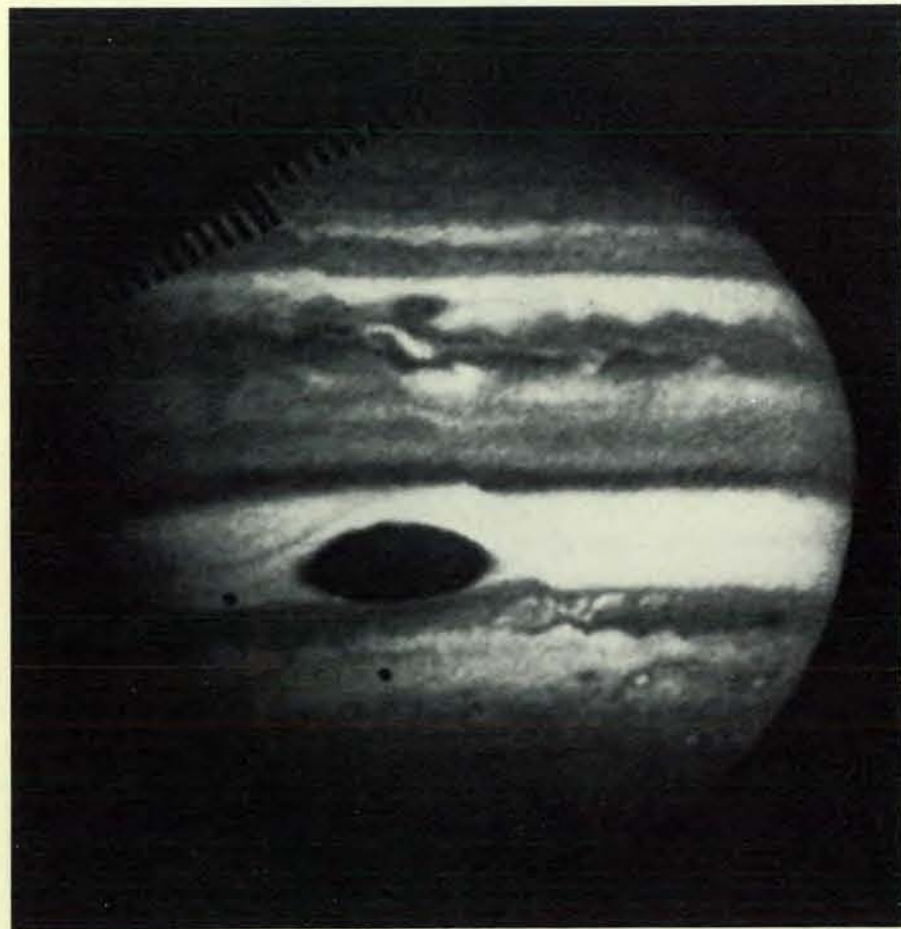
La quantità di calore irradiata da Giove indica che l'interno del pianeta è caldo; se l'interno fosse freddo non ci sarebbe stato abbastanza calore da durare fino a oggi. Una conseguenza del modello dell'interno caldo di Giove è che non vi si possono formare dei solidi. Oggi si pensa che il pianeta sia principalmente liquido, con una transizione graduale a una atmosfera gassosa nelle poche migliaia di chilometri più esterni.

Anche i dati spettroscopici delle lunghezze d'onda radio e infrarosse danno informazioni sulla pressione e la temperatura nell'atmosfera di Giove. Ai livelli più profondi la temperatura diminuisce con l'altitudine al tasso di circa due centigradi per chilometro. Questo valore è prossimo a quello della variazione adiabatica di temperatura, cioè al valore del gradiente termico in un'atmosfera ben mescolata. In un'atmosfera adiabatica piccoli volumi di gas si muovono verticalmente senza scambiare calore con quelli circostanti; diventano più caldi o più freddi soltanto perché varia la loro pressione se salgono o se discendono. Volumi adiacenti di atmosfera alla stessa quota sono indistinguibili gli uni dagli altri. Di solito un gradiente termico adiabatico indica anche che l'atmosfera viene rimescolata per convezione.

La temperatura nell'atmosfera di Giove è di 165 kelvin (gradi centigradi al di sopra dello zero assoluto) alla quota in cui la pressione è di una atmosfera (pressione dell'atmosfera terrestre al livello del mare). La temperatura seguita a diminuire con l'altezza fino a raggiungere un valore minimo di 105 kelvin alla quota in cui la pressione è 0,1 atmosfere; a questo punto ricomincia ad aumentare lentamente. Per analogia con l'atmosfera terrestre questo minimo segna l'inizio della stratosfera di Giove. In quest'ultimo strato la temperatura è controllata essenzialmente dall'irraggiamento, piuttosto che dalla convezione.

Su Giove esistono nubi sottili che giungono fino alla base della stratosfera. Dove la pressione va da 0,6 a una atmosfera cominciano a presentarsi nuvole spesse e rotte, con fori che si aprono su livelli più profondi. Quindi la pressione nella parte più alta delle nuvole è simile sulla Terra e su Giove. Invece a quei livelli su Giove la temperatura è molto più bassa, perché Giove è cinque volte più lontano dal Sole.

La sola spettroscopia non consente di stabilire la composizione delle particelle delle nuvole gioviane e la natura del materiale che le dota dei loro colori caratteristici. A questo punto sono risultati utili i calcoli teorici di John S. Lewis, del Massachusetts Institute of Technology. Lewis è partito dall'ipotesi che l'atmosfera di Giove abbia la stessa composizione di quella del Sole negli intervalli di temperatura e pressione osservati sul pianeta, e che tutti i suoi costituenti siano in equilibrio chimico. Quindi ha calcolato le quantità di materia solida e liquida nell'atmosfera in funzione della quota al di sopra di una superficie di riferimento arbitraria. I suoi calcoli mostrano che le nubi più basse e più spesse sono di acqua condensata, poiché nell'atmosfera vi è abbondanza sia d'idrogeno sia di ossigeno. Al di sopra di queste vi sono nubi di idrosolfuro di ammonio (NH_4SH), sovrastate a loro volta da nubi di ammoniaca pura (NH_3). Il livello di ogni nube dipende dalla tensione di vapore della fase che la compone, che varia fortemente con la temperatura. Le sostanze meno volatili, come l'acqua, condensano a temperature superiori (cioè a quote inferiori) di quelle più volatili come l'ammoniaca. La tensione di vapore dell'acqua, relativamente bassa alle temperature caratteristiche della sommità delle nubi di Giove, spiega anche perché il vapore acqueo vi sia stato scoperto solo da poco tempo: per spettroscopia l'acqua non può venire determinata se non è sotto forma di vapore. Nel 1974 Harold P. Larson e i suoi collaboratori hanno scoperto l'acqua su Giove per la prima volta, osservando le radiazioni provenienti dalle molecole di acqua dei livelli più profondi che passavano attraverso i fori degli strati di nubi superiori.

Il modello di Lewis, secondo il quale a ogni livello le sostanze sono in equilibrio

Raffigurazione della turbolenza dell'atmosfera gioviana in una simulazione al calcolatore preparata da Gareth P. Williams della Princeton University, il quale è ricorso a un modello che cerca di spiegare le bande est-ovest di Giove. Le linee del disegno sono linee di flusso, lungo le quali si hanno correnti atmosferiche; dove le linee sono più ravvicinate il flusso è più veloce. La direzione relativa del flusso è indicata nel grafico a destra della figura: dove la curva è a destra della linea verticale complessivamente il flusso è diretto verso est, cioè è più veloce della velocità media di rotazione del pianeta. Dove la curva è a sinistra della linea verticale complessivamente il flusso è diretto verso ovest, cioè è più lento della velocità media di rotazione. La velocità relativa del flusso è data dall'ampiezza della deviazione della curva.

chimico, non spiega i colori delle nubi. Nel suo modello le particelle di nubi sono bianche, mentre le nubi di Giove hanno delicate sfumature di rosso, marrone, bianco e blu. Tuttavia, tenendo conto che la convezione porta in alto materiale dei livelli più bassi e che nella parte superiore dell'atmosfera i raggi ultravioletti del Sole favoriscono le reazioni chimiche, non è sorprendente che parti dell'atmosfera di Giove si allontanino localmente dall'equilibrio chimico. Bastano piccole quantità di materiali coloranti a spiegare le osservazioni, e un'atmosfera con la stessa composizione di quella del Sole conterrebbe tutti gli elementi necessari in una vasta gamma di composti. In generale si pensa che le nubi siano sostanzialmente come Lewis le ha previste e che i colori si possono spiegare con l'aggiunta di piccole quantità di zolfo, fosforo e molecole organiche complesse.

I dati sulla struttura orizzontale e sui moti dell'atmosfera di Giove si hanno principalmente dalle fotografie alle lunghezze d'onda del visibile. Questi dati comprendono le osservazioni a terra eseguite a partire dalla fine del XIX secolo e le immagini ad alta risoluzione riprese dalle sonde spaziali *Pioneer 10* e *Pioneer 11*. Le osservazioni a terra prima del 1955 sono state riassunte da B.M. Peek nel classico lavoro *The Planet Jupiter*. Dall'esame di queste osservazioni è chiaro che la maggior parte delle proprietà dell'atmosfera gioviana esiste da decenni o più, e una longevità simile per le strutture delle nubi, rispetto agli standard terrestri, è notevolissima. Sulla Terra le strutture di nubi durano raramente più di una o due settimane, a meno che non siano connesse direttamente alla topografia sottostante, per esempio a catene di montagne; su Giove, per quanto se ne sa, la topografia non esiste. Per questo sarebbe assai interessante sapere perché la struttura delle nubi di Giove si conserva per tanto tempo.

Le caratteristiche dell'atmosfera di Giove, fotografate all'infrarosso (*in alto*) dalla sonda spaziale *Pioneer 10*, rivelano la distribuzione della temperatura alla sommità delle nubi. Le parti chiare sono regioni a bassa emissione infrarossa e indicano nuvole alte e spesse. Le parti scure sono regioni a elevata emissione infrarossa e indicano dei fori nello strato superiore di nubi, che scoprono gli strati più caldi sottostanti. Quando la fotografia all'infrarosso viene confrontata con una fotografia fatta alla luce visibile (*in basso*) appare evidente che le aree coperte di nubi corrispondono alle «zone» chiare e quelle prive di nuvole corrispondono alle «fasce» scure. Poiché di solito nell'atmosfera le nubi si formano nelle correnti calde ascensionali e si disperdono nelle correnti fredde discendenti, le zone devono corrispondere a regioni di gas ascensionali caldi e le fasce a gas più freddi discendenti. In basso a sinistra nelle due immagini si vede la grande macchia rossa; è un'area a emissione infrarossa particolarmente debole e assomiglia più alle zone che alle fasce. Nell'immagine alla luce visibile la forma a pettine che si vede in alto a sinistra, le due macchie nere quadrate nei pressi della grande macchia rossa e la discontinuità apparente che attraversa la fotografia sono dovute alla tecnica di ripresa.

Forse la spiegazione più semplice e più diretta della longevità delle nubi di Giove è quella avanzata da Peter J. Gierasch, che ora sta alla Cornell University, e da Richard M. Goody della Harvard University. Gierasch e Goody sottolineano che su Giove la costante di tempo d'irraggiamento è estremamente lunga. La costante di tempo d'irraggiamento è il tempo necessario a una massa d'acqua per riscaldarsi o raffreddarsi per irraggiamento nella parte infrarossa dello spettro; sulla Terra essa vale qualche settimana, cioè ha una durata paragonabile alla vita media delle strutture di flusso dell'atmosfera. Su Giove la costante di tempo d'irraggiamento è superiore a un anno. Questa differenza è una conseguenza delle temperature più basse dell'atmosfera di Giove, della quantità di calore minore che il pianeta riceve dal Sole (solo il 4 per cento di quello ricevuto dalla Terra) e del fatto che i gas dell'atmosfera gioviana emettono meno radiazioni infrarosse dei gas dell'atmosfera terrestre. Per di più se le masse di gas nell'atmosfera di Giove possono conservare calore a bassa quota, fino alla base delle nubi, allora la costante di tempo di irraggiamento può essere anche più lunga di un anno a causa del grande volume di gas che va riscaldato o raffreddato. Quindi su Giove la lunga persistenza dei fenomeni atmosferici potrebbe essere dovuta in larga parte alla lentezza con cui differenze di temperatura vengono annullate per irraggiamento.

Un altro aspetto singolare delle proprietà delle nubi gioviane è la loro organizzazione e la loro regolarità spaziale. In qualsiasi momento ci sono di solito almeno dieci bande, fasce e zone, che circondano il pianeta su linee di latitudine costante. Le fotografie dell'atmosfera terrestre fatte dai satelliti non mostrano un grado così elevato di simmetria rispetto all'asse di rotazione.

Le fasce sono marroni o marroni con sfumature blu, mentre le zone sono bianche o bianche con sfumature rosse. Le parti più rosse, come la grande macchia rossa e le macchie più piccole, in genere sono situate nelle zone. Le osservazioni all'infrarosso, che permettono di misurare la temperatura alla sommità delle nubi, dicono che c'è una differenza fondamentale fra le zone e le macchie rosse da un lato e le fasce dall'altro.

Su Giove, come sulla Terra, le differenze di temperatura più spinte entro una certa area si trovano fra uno strato e l'altro dell'atmosfera. Poiché in generale la temperatura diminuisce con l'altitudine, una temperatura alta all'infrarosso indica che ad altitudini elevate si hanno solo nubi sottili e trasparenti o che non se ne hanno affatto, mentre una temperatura bassa all'infrarosso indica che alle alte quote si hanno nuvole relativamente spesse e opache. Sulla Terra i satelliti meteorologici utilizzano questo principio per fotografare le nubi sia di giorno sia di notte con le radiazioni infrarosse: una tempesta convettiva, come un ciclone, appare sempre come una zona a bassa temperatura infrarossa, per la vasta coltre di alte nuvole che la ricopre.

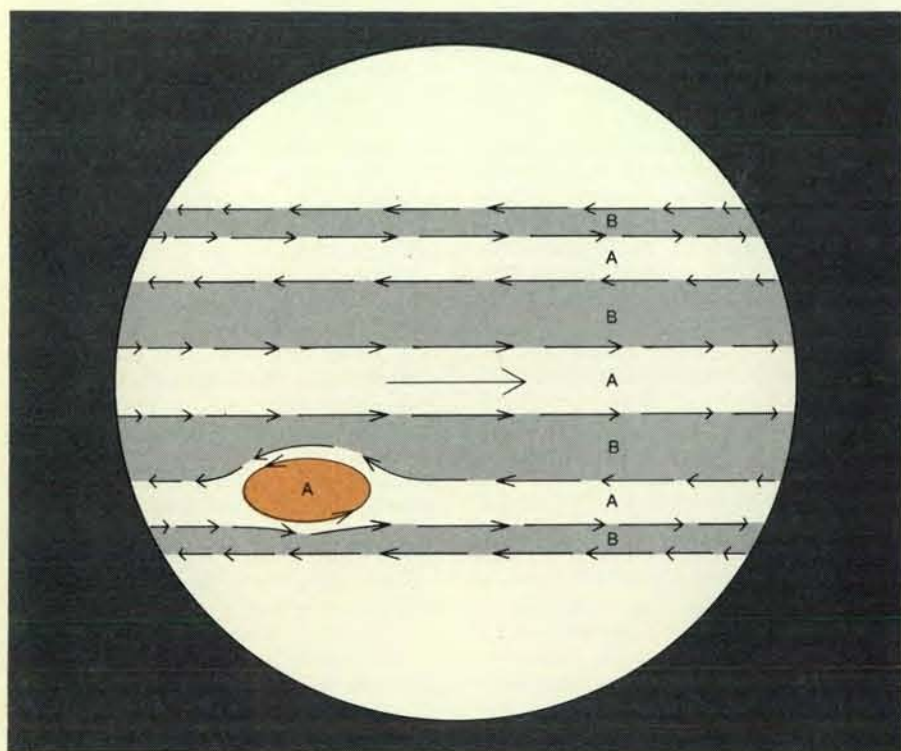
Il *Pioneer 10*, superando Giove, il 3 dicembre 1973 prese due immagini all'infrarosso; un anno più tardi il *Pioneer 11* ne prese altre quattro. Il confronto fra un'immagine all'infrarosso con una simile fatta nella zona visibile dello spettro mostra che le zone sono aree a bassa temperatura infrarossa, mentre le fasce sono aree ad alta temperatura infrarossa. La grande macchia rossa ha una temperatura infrarossa particolarmente bassa, e quindi, più che alle fasce, è simile alle zone. Da tutto ciò, per analogia con la Terra, si può ricavare che le macchie rosse e le zone bianche sono regioni di convezione attiva e di moti ascensionali dell'atmosfera, mentre le fasce di colore marrone sono regioni di dispersione di nubi e di moti discendenti.

Sebbene le osservazioni all'infrarosso siano di aiuto per classificare le caratteristiche dell'atmosfera di Giove, esse non rispondono ai principali interrogativi sul perché queste caratteristiche esistono. Le osservazioni sul moto delle nubi forniscono alcuni punti chiave per rispondere a queste domande. Dalla Terra è possibile registrare la posizione delle macchie piccole rispetto alle fasce, alle zone e ad altre strutture importanti. Le registrazioni fotografiche della superficie di Giove di cui disponiamo coprono un arco di circa 100 anni. Le macchie più piccole che si riescono a vedere dalla Terra hanno un diametro di circa 3000 chilometri, cioè circa un cinquantesimo del diametro di Giove e più o meno il diametro di un grande sistema ciclonico sulla Terra. Tali sistemi ciclonici sulla Terra si muovono secondo la direzione media del vento; per esempio nell'America settentrionale si muovono da ovest a est. A volte, tuttavia, man mano che un sistema ciclonico si evolve e si attenua, il suo moto apparente rispetto alla direzione media del vento può cambiare. Di conseguenza il ricorso alle caratteristiche delle nubi per ricavare i venti su larga scala pone alcuni problemi.

Le macchie minori, osservate per qualche settimana, si muovono sempre in senso antiorario attorno alla grande macchia rossa; attorno a essa, quindi, probabilmente i venti spirano sempre in quella direzione. Inoltre la direzione e l'intensità dei venti sono simili a quelle osservate decenni addietro nella stessa zona. Altrove su Giove i venti soffiano secondo linee di latitudine costante; le massime velocità relative si trovano ai confini fra le fasce e le zone.

Come sulla Terra, le velocità relative sono piccole rispetto alla velocità media dell'atmosfera associata alla rotazione del pianeta. Si può determinare la rotazione dell'interno di Giove in base alla rotazione del campo magnetico del pianeta, che ha un periodo di nove ore, 55 minuti e 29,7 secondi. Lo stesso periodo di rotazione è tipico per le nubi di Giove alle medie latitudini, sebbene piccole zone scartino fino a cinque minuti dal periodo medio di rotazione a causa dei venti ai bordi fra fasce e zone. In un intervallo di latitudini che si estende per circa otto gradi ai due lati dell'equatore

Profilo trasversale dell'atmosfera di Giove, fatto lungo un meridiano, per confrontare le regioni di alta e di bassa pressione. La pressione atmosferica aumenta verso il centro del pianeta. A grande profondità, dove l'atmosfera ruota uniformemente col periodo medio di rotazione del pianeta, le «superfici» atmosferiche a pressione costante (*in colore più intenso*) sono orizzontali. A profondità intermedie le maggiori temperature delle zone fanno rigonfiare verso l'alto le superfici a pressione costante (*in colore più chiaro*) creando pressioni più elevate nelle zone (*a destra*) che nelle fasce (*a sinistra*). Le forze di Coriolis generate dalla rotazione del pianeta spingono l'atmosfera a muoversi in direzione ortogonale al disegno, verso l'osservatore (verso ovest) ai bordi delle zone dal lato dell'equatore, e in direzione contraria (verso est) ai bordi delle zone dal lato dei poli. La risalita dei gas caldi in corrispondenza delle zone insieme alla discesa dei gas freddi in corrispondenza delle fasce produce nell'atmosfera una circolazione secondaria, più lenta.

L'atmosfera terrestre a scala globale è baroclina, cioè si hanno grandi differenze orizzontali di temperatura fra l'equatore e i poli. A causa di queste differenze in ambedue gli emisferi si forma nell'atmosfera un'onda longitudinale che si sposta da ovest (*a sinistra*) a est (*a destra*) e porta calore dall'equatore ai poli. Gli avvallamenti (escursioni verso l'equatore) dell'onda sono associati a regioni di bassa pressione (*B*) e le creste (escursioni verso i poli) sono associate a regioni di alta pressione (*A*). La circolazione dell'aria attorno alle regioni di bassa pressione è ciclonica.

L'atmosfera di Giove a scala globale è barotropa, cioè si hanno piccolissime differenze orizzontali di temperatura fra l'equatore e i poli. Le regioni di alta (*A*) e bassa pressione (*B*) sono strutture lineari allineate ai paralleli: le zone (*in bianco*) e le fasce (*in grigio*). Ai limiti fra le bande si hanno venti ad alta velocità; dove il moto relativo del vento è verso ovest (*verso sinistra*) i venti sono detti retrogradi. In ambedue gli emisferi, al di sopra di una latitudine di 45 gradi, la struttura a bande scompare (*si veda l'illustrazione a pagina 40*). L'ovale colorato nell'emisfero meridionale è la grande macchia rossa, che è circondata da venti anticiclonici.

gioviano, il periodo di rotazione dell'atmosfera è più breve di circa cinque minuti del periodo del campo magnetico. Lontano dalle zone equatoriali si ha il periodo di rotazione più breve, che è associato al limite fra una cospicua fascia e una zona nell'emisfero settentrionale. Là i venti compiono un giro del pianeta in 9 ore e 49 minuti, e a quel limite la velocità delle nubi rispetto all'interno rotante di Giove è di circa 120 metri al secondo, ossia più di 450 chilometri all'ora.

Quando i moti atmosferici persistono per intervalli lunghi rispetto al periodo di rotazione di un pianeta, la forza di Coriolis svolge un ruolo importante nella dinamica dell'atmosfera. Nell'emisfero settentrionale, su Giove come sulla Terra, la forza di Coriolis devia il moto da ovest verso est. Per bilanciare questa forza dev'esserci a est un'altra forza, costituita da una zona di alta pressione. Per questo sulla Terra nell'emisfero settentrionale i venti soffiano in senso antiorario, formando dei cicloni, attorno ai centri di bassa pressione, e girano in senso orario attorno ai centri di alta pressione, formando degli anticicloni. Nell'emisfero meridionale le cose vanno al contrario. Per venti che si muovono in zone a latitudine costante (come ai limiti fra le fasce e le zone di Giove) gli alti e i bassi di pressione si hanno fra le regioni a massima e minima velocità.

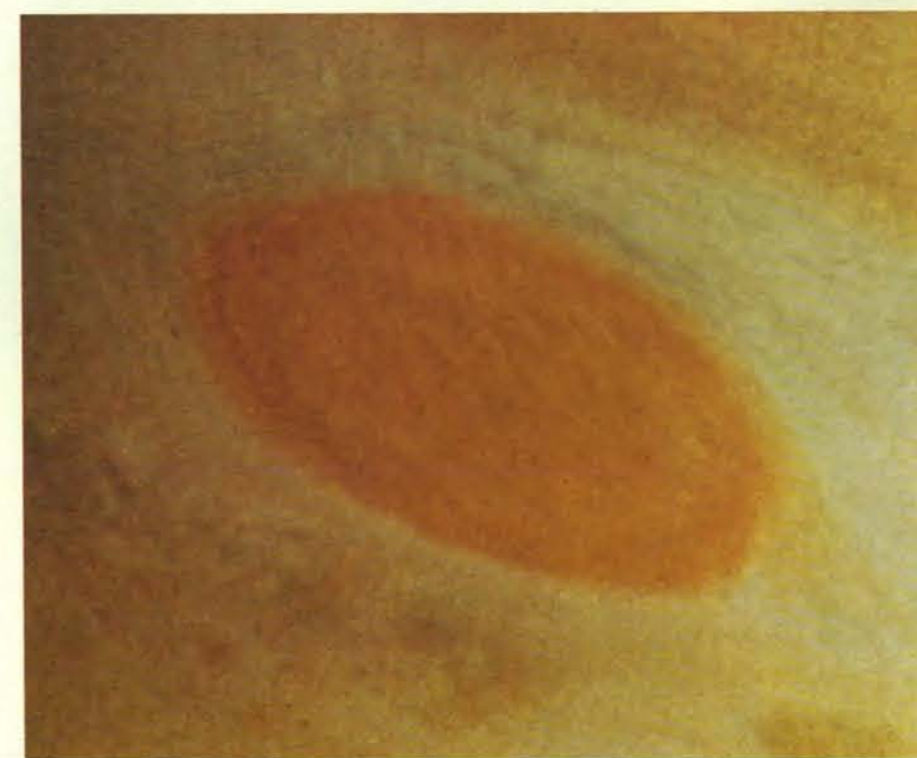
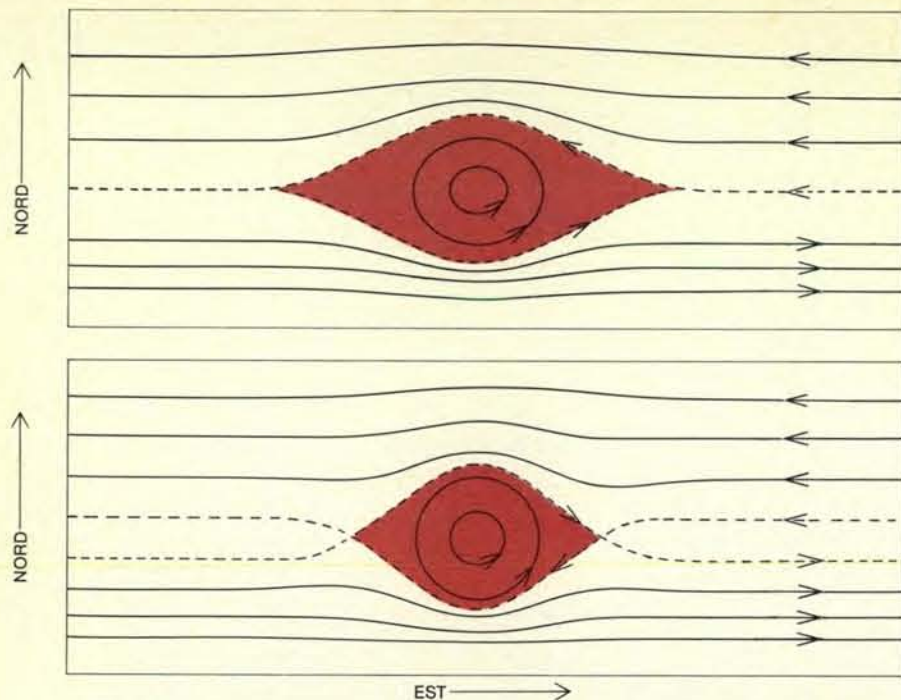
Su Giove le zone e la grande macchia rossa sono regioni ad alta pressione (anticicloniche), mentre le fasce sono regioni a bassa pressione (cicloniche). Ciò è stato messo in evidenza per la prima volta nel 1951 da Seymour L. Hess della Florida State University e da Hans A. Panofsky della New York University. Sembra quindi che le zone e la grande macchia rossa siano fondamentalmente diverse dalle tempeste terrestri, che di solito al livello del mare sono cicloniche. Tuttavia questa differenza non è profonda come potrebbe sembrare. Poiché le nubi tendono a formarsi nell'aria che risale, e poiché l'aria che risale in genere è calda, è ragionevole supporre che le zone e la grande macchia rossa sono più calde di ciò che le circonda a qualsiasi livello entro le nubi. In questo senso somigliano, sulla Terra, ai cicloni tropicali e ai cicloni extratropicali in uno stato avanzato del loro sviluppo, la maggior parte dei quali è appunto calda. La somiglianza è significativa perché sulla Terra le masse d'aria calda tendono ad avere centri di alta pressione e una circolazione anticiclonica alle alte quote, che sono appunto le quote a cui si riferiscono le osservazioni su Giove.

La ragione per cui tali tempeste sono, a quote molto alte, anticicloniche è che la caduta di pressione con l'altitudine in una massa di aria calda è minore della caduta di pressione in una massa di aria fredda. Questa è una conseguenza del rapporto idrostatico fra la variazione di pressione e la densità: se la densità è bassa e se l'aria è calda, anche la diminuzione di pressione con l'altitudine sarà bassa. Quindi una massa di aria calda, all'aumentare dell'altitudine, tende a diventare una zona di alta pressione (anticiclonica). Se a bassa quota la pressione è molto più bassa che nelle zone circostanti, come in una tempesta ciclonica terrestre, la transizione a circolazione anticiclonica può non aver luogo. Gli uragani terrestri, in generale, sono fortemente ciclonici a livello del mare e debolmente anticiclonici alle alte quote. In altri termini la circolazione anticiclonica che si osserva su Giove nelle zone e nella grande macchia rossa è coerente coi dati delle osservazioni all'infrarosso, secondo i quali si tratta di centri caldi di moti ascensionali. Perciò, in un certo senso, sono simili alle tempeste convettive calde che si hanno sulla Terra.

Però, come abbiamo visto, fra la Terra e Giove c'è una differenza fondamentale dovuta al fatto che l'atmosfera terrestre è limitata in basso da una superficie relativamente indeformabile di acqua e rocce. Questa superficie può fornire e sopportare grandi differenze di pressione atmosferica fra zone diverse al livello del mare, e così possono aversi forti venti in prossimità del suolo. Sappiamo poco sull'atmosfera di Giove al di sotto delle nubi visibili, ma è ragionevole ammettere che sotto un certo livello essa ruoti uniformemente con lo stesso periodo di rotazione del campo magnetico. Tutto ciò è equivalente all'affermazione che l'atmosfera ha un limite inferiore fluido che si deforma facilmente quando vi si applica una pressione dall'alto. Questa deformazione impedisce la formazione di differenze di pressione ai livelli inferiori e quindi mantiene debole la circolazione a quei livelli. Così, mentre sulla Terra una massa di aria calda alle basse quote può avere una circolazione ciclonica o anticiclonica, su Giove la circolazione a bassa quota dev'essere nulla. In questo senso la circolazione su Giove può somigliare di più alle correnti negli oceani che alle correnti nell'atmosfera terrestre. Le correnti oceaniche tendono a indebolirsi a grande profondità e sono sempre anticicloniche presso la superficie, se l'acqua fra i due livelli è calda.

Vi sono stati diversi tentativi di porre questi argomenti su basi quantitative. Nel 1969 Jeffrey Cuzzi, allora al California Institute of Technology, e io decidemmo di studiare le strutture dei venti zonali osservati su Giove. Per prima cosa riscoprimmo il rapporto qualitativo discusso da Hess e Panofsky nel 1951. Facemmo quindi una stima grossolana della differenza di temperatura fra zone e fasce necessaria a spiegare i venti che si osservano. In effetti la quantità da determinare non è la sola differenza di temperatura, ma il prodotto fra la differenza media di temperatura e la profondità alla quale si estende tale differenza. Se i moti zonali osservati si riferiscono alla sommità delle nubi di ammoniaca, nel modello di Lewis dell'atmosfera gioviana, allora la profondità va misurata verso il basso a partire da tale livello.

Il prodotto della differenza di tempe-

Fotografia di una macchia rossa più piccola, simile alla grande macchia rossa, fatta da *Pioneer 10*. Queste macchie, che durano circa due anni, non sono rare e di solito si trovano nelle zone. La loro somiglianza alla grande macchia rossa fa pensare che tali strutture siano fenomeni meteorologici persistenti e che non siano legati a nessuna struttura solida dell'interno di Giove.

La grande macchia rossa, qui in una fotografia del *Pioneer 11*, oggi è ritenuta simile a una tempesta tropicale terrestre. Il flusso osservato nell'atmosfera attorno alla macchia e nella zona è anticiclonico, e ciò significa che la pressione nella macchia è alta. Anche in un ciclone tropicale terrestre il flusso alla sua sommità è anticiclonico, cioè alle alte quote si ha una zona anticiclonica. La grande macchia rossa è una caratteristica della superficie di Giove da circa tre secoli. I modelli idrodinamici della macchia indicano che, in assenza di dissipazione di energia, tale struttura in linea di principio potrebbe durare per sempre. Anche una piccola dissipazione di energia non avrebbe grande effetto, perché su Giove l'energia da dissipare potrebbe derivare completamente dal calore latente ceduto dal gas ascendente nella zona, che condensa e forma le nubi.

Il modello al calcolatore della grande macchia rossa (*in alto*) dell'atmosfera di Giove riproduce diverse strutture osservate del flusso atmosferico. Esso mostra i venti in senso antiorario attorno alla macchia e la direzione normale del flusso est-ovest a nord e a sud della macchia. Mostra anche le estremità a punta dei bordi orientali e occidentali della macchia. Una variazione del modello (*in basso*) indica in che modo le macchie scure più piccole, che talvolta si avvicinano alla grande macchia rossa, possano a volte cambiare improvvisamente di latitudine e cominciare a retrocedere, spinte all'indietro da una corrente di gas. In ambedue le versioni del modello la scala orizzontale del diagramma è stata compressa di un fattore tre rispetto alla scala verticale.

ratura per la profondità, sulla base dei venti osservati su Giove, è circa 150. In altri termini la profondità reale oltre la quale le differenze di temperatura fra zone e fasce diventa apprezzabile è sconosciuta, ma se questa profondità fosse di 15, 150 o 1500 chilometri la differenza di temperatura a quella profondità sarebbe rispettivamente di 10, 1 o 0,1 gradi centigradi, per poter spiegare i venti che si osservano. La grandezza della differenza di temperatura dipende dalle sorgenti e dai pozzi di calore che mantengono attiva la circolazione atmosferica. La profondità critica è lo spessore della regione che fa da sorgente e pozzo di calore. Per discutere ulteriormente questo problema è necessario considerare il meccanismo sorgente-pozzo.

Nel 1971 Gierasch, che allora era alla Florida State University, e Albert I. Barcilon hanno avanzato l'ipotesi che l'atmosfera di Giove sia simile all'atmosfera della Terra al di sopra dei tropici, dove il Sole riscalda la superficie dell'oceano causando l'evaporazione dell'acqua. L'aria umida vicino alla superficie diventa instabile e dà inizio alla convezione e alla formazione di nubi cumuliformi. Sui tropici si hanno moti organizzati e su larga scala, perché la convezione dei cumuli su piccola scala, che è il mezzo principale con cui si trasferisce calore dall'oceano all'atmosfera, varia a seconda dei moti su larga scala. Ai tropici le massime differenze orizzontali di temperatura fra un posto e l'altro si hanno fra l'aria che è stata riscaldata per condensazione e quella che non lo è stata. Secondo Gierasch e Barcilon la differenza di temperatura fra le zone e le fasce di Giove, come quelle che si hanno sui tropici terrestri, è dovuta alla liberazione di calore latente.

Supponiamo che il modello di Lewis per le nubi si adatti alle zone gioviane, e che le fasce siano secche, cioè prive di vapori condensabili; in questo caso le zone hanno una temperatura superiore di circa due gradi a quella delle fasce, a qualsiasi livello superiore a quello in cui condensa il vapore acqueo. Di conseguenza, una differenza di temperatura di circa due gradi si estende per una profondità di circa 75 chilometri, cioè per tutto lo spessore del sistema di nubi. Inoltre il prodotto della differenza di temperatura per la profondità è in accordo con il valore di 150 dedotto dalle osservazioni del vento.

Gierasch e Barcilon non hanno spiegato il fatto che i vapori condensabili restano concentrati nelle zone. Sulla Terra, in un ciclone tropicale, i venti ciclonici esercitano uno *stress* ciclonico (forza per unità di area) sulla superficie oceanica. Questo stress, interagendo con la forza di Coriolis, spinge l'aria verso il centro a bassa pressione del ciclone, appena al di sopra della superficie, mentre l'acqua, appena sotto la superficie, viene spinta in fuori. Sulla Terra il flusso nell'oceano si può trascurare, ma il convergere di aria umida verso il centro del ciclone rifornisce il ciclone di calore latente mantenendolo così in circolazione.

Su Giove le regioni nuvolose sono anticicloniche e ruotano in senso opposto a quello delle regioni nuvolose terrestri. Non c'è oceano; il fluido inferiore è semplicemente l'atmosfera sottostante le nubi. Uno stress anticiclonico, esercitato dall'alto sul fluido inferiore, spinge questo fluido verso l'interno e l'atmosfera superiore verso l'esterno. In questo caso, quello che ci interessa è il fluido inferiore che converge verso il centro. La convergenza avviene sotto la base delle nubi, ai livelli in cui le precipitazioni evaporano; è questa convergenza che rifornisce continuamente le zone di vapori condensabili. Questo processo è uno dei possibili meccanismi per spiegare come si conserva la differenza di temperatura fra zone e fasce.

Non ho detto nulla, finora, sulla circolazione dell'atmosfera alle medie latitudini su Giove e sulla Terra. Sulla Terra la sorgente principale di energia per l'atmosfera alle medie latitudini è il gradiente termico orizzontale, che ha origine dal fatto che il Sole riscalda l'equatore molto più dei poli. Un'atmosfera con tali differenze di temperatura ad altitudine costante viene detta baroclina, e in genere è instabile. L'instabilità si manifesta nel flusso principale dell'atmosfera verso est come una struttura a onda dove ogni avvallamento rappresenta il punto in cui l'aria compie la massima escursione verso l'equatore e ciascuna cresta il punto in cui compie la massima escursione verso i poli. L'onda porta calore sia nella parte superiore dell'atmosfera sia verso i poli. Il flusso di calore verso l'alto, in cui l'aria calda risale e quella fredda discende, tende a abbassare il centro di massa dell'atmosfera, perché un certo volume di aria fredda è più denso e quindi più pesante dello stesso volume di aria calda; il trasporto di calore libera, dunque, energia potenziale gravitazionale. Gli avvallamenti e le creste dell'onda baroclina sono rispettivamente i cicloni e gli anticicloni che caratterizzano il clima terrestre alle medie latitudini.

Un problema importante è se queste onde barocline siano presenti su Giove. Se il flusso di calore dall'interno di Giove fosse uniforme dall'equatore ai poli, il calore totale ceduto all'atmosfera sia dall'interno del pianeta sia dal Sole sarebbe all'equatore sempre il doppio che ai poli, perché queste due regioni vengono riscaldate dal Sole in modo diverso. Questa differenza potrebbe causare una forte instabilità baroclina, se alle medie latitudini Giove fosse simile alla Terra.

Tuttavia il fatto che Giove si presenta a bande e simmetrico rispetto a un asse indica che a quelle latitudini Giove deve essere notevolmente diverso dalla Terra. Un aspetto essenziale dell'instabilità baroclina è il mescolamento fra regioni a diversa latitudine, e sebbene su Giove vi siano alcuni disturbi che somigliano alle onde barocline, essi sono confinati in sottili bande di latitudine e non sembrano dare luogo a mescolamenti su vasta scala fra una banda e l'altra. Inoltre se il trasporto di calore verso i poli per instabilità baroclina fosse l'unico modo di

STRUTTURA	MISURE ALL'INFRAROSSO	ALTEZZA DELLE NUBI	CIRCOLAZIONE	PRESSIONE	TEMPE-RATURA	VELOCITA' VERTICALE	NUBI PREVISTE	COLORE
FASCIA	CALDA	PICCOLA	CICLONICA	BASSA	BASSA	IN GIÙ	BASSE, SOTTILI	SCURO
ZONA	FREDDA	GRANDE	ANTICICLONICA	ALTA	ALTA	IN SU	ALTE, SPESSE	CHIARO
GRANDE MACCHIA ROSSA	FREDDA	GRANDE	ANTICICLONICA	ALTA	ALTA	IN SU	ALTE, SPESSE	ARAN-CIONE

Riassunto delle strutture caratteristiche dell'atmosfera di Giove. La grande macchia rossa è simile alle zone da tutti i punti di vista importanti salvo la forma, mentre da tutti i punti di vista le fasce sono il contrario delle zone. Nel complesso questi dati fanno pensare che queste strutture non siano fenomeni isolati e distinti, ma debbano essere collegati assieme come parti di una struttura atmosferica globale.

compensare il calore in più che giunge all'equatore di Giove, fra l'equatore e i poli si svilupperebbe una notevole differenza di temperatura.

Peter H. Stone, del MIT, ha una teoria generale sull'instabilità baroclina in qualsiasi atmosfera planetaria. Per Giove egli ha valutato che se il calore solare fosse trasportato ai poli solo dall'instabilità baroclina l'equatore avrebbe una temperatura di 30 gradi superiore a quella dei poli. Recentemente la sua stima è stata controllata, quando i rivelatori all'infrarosso del *Pioneer 10* e del *Pioneer 11* hanno eseguito le prime riuscite misurazioni di calore ai poli di Giove.

Queste misure dimostrano che il calore emesso dai poli e dall'equatore di Giove è circa lo stesso, e questo implica che la temperatura dei poli sia più o meno uguale a quella dell'equatore agli stessi livelli di pressione nell'atmosfera. In effetti la differenza di temperatura osservata non è superiore a tre gradi e ciò comporta che Giove, alle medie latitudini, sia diverso dalla Terra e che nell'atmosfera di Giove il trasporto di calore per instabilità baroclina sia insignificante.

Se le cose stanno in questi termini, la differenza di calore solare assorbito da Giove all'equatore e ai poli dev'essere compensata da altro calore proveniente dall'interno del pianeta. Recentemente ho proposto un meccanismo in cui questa differenza potrebbe venire annullata da moti convettivi nell'interno del pianeta. Questo meccanismo funzionerebbe da termostato: quando il valore effettivo del gradiente termico nell'atmosfera è uguale al gradiente adiabatico, il flusso di calore per convezione è nullo, mentre quando la diminuzione di temperatura con l'altitudine è un po' più rapida di quella adiabatica il flusso di calore per convezione è grande. Così un piccolo raffreddamento dei poli rispetto all'equatore, dovuto alla differenza di energia solare assorbita, provoca ai poli un grande aumento del flusso di calore per convezione rispetto a quello dell'equatore. Il termostato convettivo assicura che tutto Giove sia quasi perfettamente adiabatico, e quindi i rapporti fra pressione e temperatura all'equatore e ai poli devono essere quasi uguali. La possibilità che su Giove vi sia un flusso di calore per convezione maggiore ai poli sembra confermata dal fatto che a circa 45 gradi di latitudine nord compaiono strutture simili a quelle che dà la convezione.

La sorgente interna di calore potrebbe anche fornire energia alla circolazione che si osserva nelle fasce e nelle zone di Giove, ed è stato ipotizzato che fasce e zone siano soltanto le manifestazioni superficiali di gigantesche celle convettive, che potrebbero avere una profondità dello stesso ordine della larghezza e che quindi potrebbero spingersi giù verso il centro del pianeta per una frazione considerevole del suo raggio.

Contro l'ipotesi delle celle di convezione ci sono diversi argomenti, ma nessuno è conclusivo. Primo, se l'atmosfera di Giove è davvero analoga all'atmosfera tropicale terrestre non ha bisogno di una circolazione profonda su larga scala sotto le nubi. I vapori condensabili potrebbero essere concentrati nelle zone e dispersi nelle fasce in uno strato superficiale proprio sotto la base delle nubi; le differenze di temperatura che hanno origine dalla condensazione nelle nubi sembrano in grado di spiegare pienamente i venti osservati. Quindi le fasce e le zone potrebbero essere fenomeni superficiali, che si estenderebbero solo un poco più in basso della base delle nubi più profonde.

Secondo, se le fasce e le zone fossero associate a una convezione termica interna a grande scala, le emissioni infrarosse di Giove dovrebbero essere più intense nelle zone, perché le zone sono sede dei moti ascensionali. Ora è vero che noi possiamo osservare il flusso di calore solo nella parte superiore dell'atmosfera, ma in ogni caso le zone non sembrano le regioni principali di emissione infrarossa. Sono invece le fasce che emettono la maggior parte dell'energia infrarossa, e che anche assorbono la maggior parte dell'energia solare. D'altra parte nelle zone e nelle fasce il flusso netto, cioè la differenza fra l'energia emessa e quella assorbita, è circa lo stesso. Questa uguaglianza significa che il flusso di calore interno alla base delle nubi è lo stesso nelle fasce e nelle zone, dato che entro le nubi non si ha trasporto di calore in senso longitudinale. Il punto è che i dati sul flusso radiativo non forniscono prove a sostegno dell'ipotesi delle celle convettive, anche se non la escludono.

Terzo, l'ipotesi delle celle convettive presenta una parte più che abbondante di difficoltà teoriche. Tipicamente la convezione è un fenomeno a piccola scala. Nel Sole o nell'atmosfera terrestre l'estensione longitudinale delle principali celle convettive è circa uguale alla distanza verticale in cui si ha un aumento di densità e di pressione di un fattore *e*, pari a 2,718. Su Giove l'altezza della scala è qualche decina di chilometri, mentre le fasce e le zone si estendono per circa 10 000 chilometri. Inoltre le strutture convettive di un corpo in rotazione tendono a svilupparsi in senso longitudinale, esattamente il contrario di quello che succede su Giove. Questi ragionamenti fanno pensare che il modo principale in cui il calore viene scambiato sia la convezione su piccola scala.

Tuttavia ci sono altri modi in cui le fasce e le zone di Giove potrebbero far parte di una profonda struttura convettiva, estesa per una frazione significativa del raggio del pianeta. Diversi studi, teorici e di laboratorio, hanno dimostrato che in una sfera rotante la convezione prende la forma di colonne lunghe e sottili, i cui assi sono paralleli all'asse di rotazione; le estremità di ciascuna colonna intersecano la superficie visibile della sfera a latitudini opposte negli emisferi settentrionale e meridionale. Friedrich H. Busse, della Università della California a Los Angeles, ha avanzato l'ipotesi che le fasce e le zone siano le manifestazioni superficiali di queste colonne lunghe e sottili; l'esistenza di queste ultime, però, è stata finora dimostrata solo in esperimenti di laboratorio eseguiti con liquidi relativamente viscosi, e quindi estendere tale ipotesi fino a comprendere gas comprimibili su una sfera rotante delle dimensioni di Giove è discutibile.

Sia che le fasce e le zone siano fenomeni superficiali limitati agli strati di nubi, sia che facciano parte di un profondo sistema convettivo, non si ha ancora una spiegazione soddisfacente del perché esse sono orientate in direzione est-ovest. Anche l'atmosfera tropicale terrestre è a bande, con una zona nuvolosa attorno all'equatore e due fasce limpide verso nord e verso sud. La zona di convergenza equatoriale intertropicale è una banda di nubi cumuliformi e di cicloni che di solito occupa una striscia fra cinque e 10 gradi a nord dell'equatore; complessivamente è una regione di risalita dell'atmosfera, anche se il calore viene trasportato verso l'alto fino alla base della stratosfera da un sistema di correnti sia ascendenti sia discendenti. Ai due lati della zona di convergenza intertropicale si hanno due fasce secche, che si estendono fino a 30 gradi a nord e a sud dell'equatore, in cui i movimenti sono quasi sempre discendenti; i moti ascendenti sono ristretti in un sottile strato prossimo al suolo.

Di solito il cielo è sereno e il calore viene scambiato verticalmente soprattutto per irraggiamento infrarosso. Lo spostamento a nord della zona di convergenza intertropicale sembra dovuto al fatto che negli emisferi settentrionale e meridionale gli oceani e i continenti sono diversamente distribuiti.

La differenza fra Giove e la Terra viene fuori a latitudini maggiori. Su Giove il sistema di bande si ripete ancora parecchie volte, mentre sulla Terra viene cancellato dalle instabilità barocline; la differenza può essere dovuta all'effetto termostatico di una sorgente interna di calore di Giove. Sulla Terra le instabilità barocline sono l'unico modo di bilanciare il calore solare nell'atmosfera: all'equatore l'aria si riscalda rispetto a quella ai poli, si sviluppano delle instabilità e ben presto la circolazione alle medie latitudini è dominata dai cicloni e dagli anticicloni. Le regioni equatoriali di Giove si riscaldano di poco rispetto a quelle polari, e il flusso di calore interno si ridistribuisce in modo tale da bilanciare la differenza di riscaldamento solare.

Il problema dell'instabilità delle fasce e delle zone è interessante per se stesso. Ogni stato di equilibrio di un'atmosfera, salvo quello di rotazione uniforme, possiede energia cinetica e energia potenziale gravitazionale. Una perturbazione di piccola ampiezza può essere in grado di estrarre parte di questa energia e crescere a spese dello stato fondamentale; quest'ultimo è stabile soltanto se tutte le perturbazioni rimangono piccole, ed è instabile se c'è almeno un tipo di perturbazione che seguita a crescere. In tutte le situazioni reali sono sempre presenti in una certa misura tutti i tipi di perturbazione, e quindi uno stato fondamentale instabile si disperde o si evolve sempre in un tempo finito.

Lo stato fondamentale più semplice è quello di un'atmosfera che scorre esattamente da ovest a est, in cui la velocità del vento varia solo con la latitudine. Questo potrebbe essere il caso di Giove, anche se potrebbe essere scorretto trascurare le variazioni della velocità con l'altitudine. Nello stato fondamentale l'unica energia disponibile è cinetica ed è associata alle diverse velocità alle diverse latitudini. Allo stesso livello fra una zona e l'altra non si hanno gradienti termici orizzontali. Si dice che questa atmosfera è barotropa, a differenza di quella baroclina, in cui l'energia è prevalentemente gravitazionale ed è associata a gradienti termici orizzontali.

Le diverse velocità del vento diretto a est possono essere messe in diagramma rispetto alla latitudine per costruire un profilo di velocità. In un'atmosfera barotropa è possibbile un'instabilità solo quando la curvatura del profilo di velocità supera un certo valore critico, uguale a due volte il coseno della latitudine moltiplicato per la velocità di rotazione del pianeta e diviso per il suo raggio. Nel 1969 Cuzzi e io applicammo il criterio di stabilità ai dati a lungo termine di Peek sul periodo di rotazione di Giove rispetto alla latitudine. Trovammo che secondo questo criterio su Giove vi sono soltanto alcune latitudini alle quali sono possibili instabilità anche marginali. Le instabilità si trovano al centro delle veloci correnti retrograde dell'atmosfera di Giove, alle latitudini in cui il periodo di rotazione è massimo. Le perturbazioni che si sviluppano più rapidamente sono onde che oscillano in senso longitudinale; in prossimità delle correnti retrograde, quindi, ogni volta che il parametro di stabilità viene superato, il flusso principale est-ovest dovrebbe disperdersi in un sistema di onde. Quando il parametro di stabilità non viene superato la struttura del flusso est-ovest resta costante nel tempo.

Il *Pioneer 11*, coi suoi primi piani dell'emisfero settentrionale di Giove, ha fornito una conferma qualitativa. Le immagini mostrano diverse bande in cui un sistema di onde fa pensare che vi sia localizzata un'instabilità. I sistemi di onde più notevoli si hanno ai bordi delle zone verso l'equatore (ossia ai bordi delle fasce verso i poli). Secondo i dati di Peek le correnti retrograde sono localizzate nelle stesse posizioni. In altre parole sembra che le instabilità si sviluppino proprio dove è previsto dal criterio di stabilità barotropa.

Dalle immagini riprese dal *Pioneer 11* sembra anche che il flusso parallelo delle fasce e delle zone, a latitudini superiori a circa 45 gradi, si dissolva completamente. Almeno qualitativamente ciò è in accordo col criterio di stabilità barotropa, perché il valore critico della curvatura è proporzionale al coseno della latitudine, il quale ai poli diventa zero; secondo il criterio di stabilità, quindi, ai poli una struttura a bande est-ovest sarebbe instabile e non potrebbe essere una caratteristica permanente dell'atmosfera.

Una struttura nord-sud è diversa. La struttura nord-sud più generale è un misto di onde che si propagano verso ovest a varie velocità. La differenza fra le strutture di flusso nord-sud e quelle est-ovest potrebbe spiegare i risultati ottenuti recentemente al calcolatore da Gareth P. Williams del Laboratorio geofisico di dinamica dei fluidi della Princeton University. Egli è partito dal modello barotropo dell'atmosfera di Giove e ha introdotto nell'equazione un termine che descriveva una forza variabile casualmente e un termine che descriveva la dissipazione di energia attribuibile alla viscosità dell'atmosfera. Ha trovato che inizialmente nel suo modello comparivano sia moti est-ovest sia moti nord-sud. Successivamente i moti nord-sud continuavano a formarsi e a scomparire, mentre la struttura est-ovest cresceva lentamente fino al valore limite imposto dal criterio di stabilità. Calibrando i rapporti fra il termine della forza variabile casualmente e il termine di dissipazione, Williams è stato in grado di costruire al calcolatore un modello di Giove dall'apparenza molto realistica (*si veda l'illustrazione a pagina 31*).

È ancora difficile capire come mai il modello barotropo riesca così bene a spiegare le proprietà dell'atmosfera di Giove. L'atmosfera profonda deve ruotare solidale con l'interno del pianeta, e quindi la velocità delle correnti ai limiti fra fasce e zone deve variare con la profondità; associate a queste variazioni con la profondità ci devono essere differenze di temperatura fra fasce e zone su piani orizzontali. In altri termini l'atmosfera di Giove dovrebbe essere baroclina e non barotropa. Il modello barotropo sarebbe ancora un'approssimazione valida se la diminuzione di temperatura effettiva fosse molto più bassa della diminuzione adiabatica o se la temperatura aumentasse con l'altitudine; ma queste sono possibilità improbabili, perché sia la teoria sia le osservazioni dicono che il gradiente termico reale è prossimo a quello adia-

Struttura verticale delle nubi dell'atmosfera gioviana, calcolata in base a un modello teorico messo a punto da John S. Lewis del Massachusetts Institute of Technology. In un'atmosfera composta principalmente di idrogeno e elio, che sono gas non condensabili, si hanno diversi strati di nubi. La linea in colore indica la temperatura dell'atmosfera alle varie profondità e pressioni e questi valori si basano sull'analisi dei dati all'infrarosso eseguita da Glenn S. Orton del Jet Propulsion Laboratory del California Institute of Technology. La linea nera tratteggiata indica la temperatura teorica dell'atmosfera alle stesse profondità e pressioni nel caso che l'atmosfera fosse perfettamente adiabatica (ossia completamente mescolata per convezione). L'atmosfera del pianeta è quasi adiabatica, salvo che ai livelli più alti.

Il bilancio dell'irraggiamento nell'atmosfera della Terra (*in alto*) e di Giove (*in basso*) mostra differenze sorprendenti fra i due pianeti. In tutti e due i pianeti la quantità di radiazioni termiche emesse (*in colore*) è quasi indipendente dalla latitudine, mentre la quantità di luce solare assorbita (*in nero*) dipende fortemente dall'angolo d'incidenza del Sole ai poli. La Terra ha una sorgente interna di calore trascurabile, e quindi per la Terra le aree delimitate dalle due curve sono uguali. La quantità in eccesso del calore solare assorbito all'equatore viene trasportata verso i poli da correnti atmosferiche e oceaniche. Giove possiede una notevole sorgente interna di calore e quindi per Giove l'area delimitata dalla curva in colore è 1,9 volte più grande di quella delimitata dalla curva in nero; inoltre su Giove pochissimo calore viene trasportato ai poli da correnti atmosferiche. Perciò il calore interno di Giove deve raggiungere la superficie preferenzialmente ai poli. Le irregolarità nelle curve di Giove sono dovute alle fasce e alle zone; le fasce emettono radiazione termica e assorbono luce solare in misura maggiore delle zone. Per ambedue i pianeti le curve sono mediate rispetto alla longitudine e al tempo diurno e stagionale.

batico, fatta eccezione per la stratosfera, che è al di sopra della zona dove si verificano i moti delle nubi. Forse il successo del modello barotropo è dovuto ai nostri tentativi eccessivamente zelanti volti a far concordare i dati delle osservazioni con una teoria basata sull'esperienza terrestre; il meccanismo reale in atto su Giove potrebbe in realtà ancora sfuggirci.

Torniamo ora alla grande macchia rossa e a macchie simili dell'atmosfera di Giove. Per diversi anni l'unica teoria in grado di spiegare la grande macchia rossa sembrava che fosse l'ipotesi della colonna di Taylor, avanzata da Raymond Hide, che allora era al MIT. A causa della lunga persistenza della grande macchia rossa, della sua costanza in latitudine e della sua unicità, sembrava che dovesse essere legata a un oggetto solido sottostante o a una caratteristica topografica che desse origine direttamente alle strutture di flusso visibili alla superficie. La colonna di Taylor è il cilindro di fluido stagnante che si pensava collegasse questo oggetto solido alla nuvola rossa che si vede nella parte superiore dell'atmosfera di Giove.

Poiché oggi si pensa che Giove non possieda superficie solida a nessuna profondità, l'ipotesi della colonna di Taylor non è più obbligatoria. Inoltre le osservazioni più recenti hanno messo in evidenza le somiglianze fra la grande macchia rossa e le zone: in effetti la lunga persistenza della macchia e la sua costanza in latitudine non sono più singolari della lunga persistenza e della costanza in latitudine della zona in cui si trova. Anche la grande macchia rossa si sposta irregolarmente verso ovest rispetto al campo magnetico di Giove, che presumibilmente sarebbe solidale con la superficie solida del pianeta, se esistesse. Infine anche altre zone sembrano possedere le loro macchie rosse, e ciò fa pensare che la grande macchia rossa non sia unica.

L'opinione che prevale oggi è che tutte le macchie rosse, così come le zone, siano fenomeni meteorologici. Poiché la grande macchia rossa ha nubi più alte e una circolazione anticiclonica più intensa di tutte le altre zone, potrebbe venire chiamata una superzona: se le zone e le macchie rosse sono generate dal calore latente di condensazione, allora potrebbe anche trattarsi di un uragano gioviano.

Come in una tempesta tropicale terrestre, c'è, nella grande macchia rossa, la probabilità di un movimento complessivo verso l'alto e, nelle zone immediatamente esterne, quella di un moto discendente relativamente uniforme. Comunque le velocità di gran lunga più elevate sono associate al flusso orizzontale attorno alla macchia; il moto verticale e la conseguente espansione dell'atmosfera alla sommità delle nubi sono troppo piccoli per poter essere misurati. Forse gli scambi verticali di energia e di calore, e così anche la dissipazione di energia, sono piccoli: quindi, in prima approssimazione si possono trascurare tutti i processi di dissipazione e ci si può chiedere se c'è un modello adeguato del flusso orizzontale che si ha attorno alla macchia. Ancora una volta il caso più semplice è quello di un'atmosfera barotropa. Lo scopo di tale studio sarebbe quello di vedere se ci sono condizioni particolari in cui le macchie rosse possono o non possono esistere e di vedere se il modello può prevedere qualche dettaglio della struttura di flusso di tali macchie paragonabile alle osservazioni.

La struttura a bande scompare alle alte latitudini su Giove, come si vede in questa fotografia ad alta risoluzione dell'emisfero settentrionale del pianeta, fatta da *Pioneer 11*. L'equatore è la seconda zona chiara a partire dal basso. La regione disturbata a nord cade alla latitudine in cui in base alla teoria il flusso fondamentale fra una fascia e una zona diventerebbe instabile. Più a nord tale instabilità è dominante a tutte le latitudini. I bordi della fotografia sono distorti per la rotazione del pianeta e per il moto della sonda spaziale.

La grande macchia rossa giace in una zona dell'emisfero meridionale di Giove e ruota in senso antiorario, come una ruota fra due superfici che si muovono in senso opposto. A nord (verso l'equatore) sulla macchia, lungo il margine settentrionale della zona, il flusso è verso ovest, mentre a sud (verso il polo) di essa il flusso è verso est. Nel 1970 lavorando con un modello elaborato al calcolatore, ho trovato che la configurazione a ruota era l'unica che dava una configurazione stazionaria (indipendente dal tempo) del flusso. Questa configurazione mostra che le macchie anticicloniche, come la grande macchia rossa, possono esistere solo in strutture anticicloniche lineari, come le zone. Questo aspetto del modello al calcolatore è confermato dal fatto che le macchie rosse si trovano associate alle zone sia nell'emisfero settentrionale sia in quello meridionale. Il modello al calcolatore prevede anche che la grande macchia rossa dovrebbe avere ai bordi orientali e occidentali, dove le opposte correnti di vento della zona si dividono, delle estremità a punta. Questo aspetto è confermato dai primi piani della macchia, in cui le estremità a punta sono chiaramente visibili. Recentemente Tony Maxworthy e Larry G. Redekopp, della University of Southern California, hanno riprodotto molti altri dettagli del flusso attorno alla grande macchia rossa. Si suppone che il flusso fondamentale sia barotropo, e così questo modello condivide la debolezza dei modelli barotropi: il gradiente termico effettivo deve differire dal gradiente adiabatico in misura incompatibile con gli altri dati sull'atmosfera di Giove.

Tutti gli studi sulla grande macchia rossa dimostrano che tali strutture di flusso possono esistere in uno stato stazionario senza una superficie solida che ne definisca la forma: in mancanza di dissipazioni di energia queste strutture potrebbero durare per sempre. Anche introducendo una piccola dissipazione, cosa appropriata per Giove, la situazione non cambierebbe apprezzabilmente: l'energia che dev'essere dissipata potrebbe essere fornita completamente dal calore latente ceduto quando i gas asciutti convergono con quelli umidi.

Un giorno o l'altro mi piacerebbe elaborare al calcolatore un modello che funzioni sia per Giove sia per la Terra. Fissando le sorgenti e i pozzi di energia appropriati, con condizioni limite più ristrette, con ipotesi opportune sulle sorgenti interne di calore di Giove e così via, il modello dovrebbe poter prevedere realisticamente il comportamento delle atmosfere di ambedue i pianeti. Avere un modello di questo tipo significherebbe comprendere la meteorologia terrestre a un livello assai più profondo, perché l'universalità dei principi che la riguardano sarebbe stata verificata dalla applicabilità a due sistemi atmosferici indipendenti.

Cesserà l'espansione dell'universo?

La recessione delle galassie lontane, la densità media della materia, l'età degli elementi chimici e l'abbondanza di deuterio suggeriscono concordemente che l'espansione non può essere arrestata né invertita

di J. Richard Gott III, James E. Gunn, David N. Schramm e Beatrice M. Tinsley

La cosmologia è campo di indagine di antica origine, ma solo negli ultimi 50 anni circa si è cominciato a capire come è nato l'universo e quale può essere il suo destino finale. Un passo decisivo fu compiuto negli anni venti, quando Edwin P. Hubble dimostrò che le nebulose a spirale non sono oggetti locali ma sistemi indipendenti di stelle molto simili al nostro, provando così che l'universo ha dimensioni molto maggiori di quanto si credesse. Hubble mostrò inoltre che l'intero sistema osservabile di galassie è in moto ordinato. Come è oggi ben noto, tale movimento ha carattere di espansione: tutte le galassie distanti si allontanano da noi.

Il fatto che l'universo sia in espansione è ritenuto oggi una certezza. Rimane invece aperto il problema se l'espansione continuerà per sempre oppure se le galassie, che ora si allontanano, si arresteranno un giorno e invertiranno il loro moto, cadendo alla fine l'una sull'altra in un immane collasso. Risolvere questo problema significa determinare quale sia la geometria dell'universo, cioè la natura dello spazio e del tempo. Se l'espansione procederà indefinitamente, allora l'universo è «aperto» e infinito; se un giorno avrà termine e invertirà la sua direzione, allora l'universo è «chiuso» e di estensione finita.

Per decidere tra queste possibilità gli astronomi costruiscono modelli matematici dell'universo e cercano poi di trovare caratteristiche osservabili dell'universo fisico che siano capaci di confermare uno dei modelli e di escludere tutti gli altri. Finora non è stata compiuta nessuna misura sperimentale singola sufficientemente precisa da risolvere il problema inequivocabilmente. Tuttavia sono possibili numerosi test indipendenti, e varie tessere del mosaico sono state fornite da molti ricercatori facendo uso di tecniche diverse. Sembra possibile, allo stadio attuale a cui è giunta la ricerca, riunire in un quadro unitario i vari pezzi. Nel loro complesso i dati che sono disponibili suggeriscono che l'universo è aperto e che la sua espansione non si arresterà mai.

Espansione isotropa

Hubble scoprì il moto di recessione delle galassie lontane mediante misure compiute sui loro spettri ottici. Gli spettri della maggior parte delle stelle (e quindi delle galassie) sono interrotti da righe scure che rappresentano l'assorbimento di una particolare lunghezza d'onda da parte degli atomi negli strati più esterni e più freddi dell'atmosfera stellare; ogni elemento chimico dà luogo a uno schema caratteristico di righe le cui lunghezze di onda sono state determinate con grande precisione in esperimenti di laboratorio. Quando la galassia si allontana dall'osservatore, la lunghezza d'onda di ogni riga spettrale aumenta in seguito all'effetto Doppler, così che tutte le righe appaiono spostate verso lunghezze d'onda maggiori e, in particolare, verso la regione rossa della banda visibile dello spettro. Questo fenomeno è noto come spostamento verso il rosso ed è possibile, misurandone l'entità, risalire alla velocità di recessione. Quando un oggetto si avvicina all'osservatore, le lunghezze d'onda delle sue righe spettrali diminuiscono per l'effetto Doppler e le righe appaiono spostate verso la regione blu dello spettro. Questo effetto è detto spostamento verso il blu. Tutte le galassie lontane di cui Hubble e altri dopo di lui hanno misurato gli spettri presentano spostamenti verso il rosso: si deduce pertanto che esse stanno allontanandosi da noi.

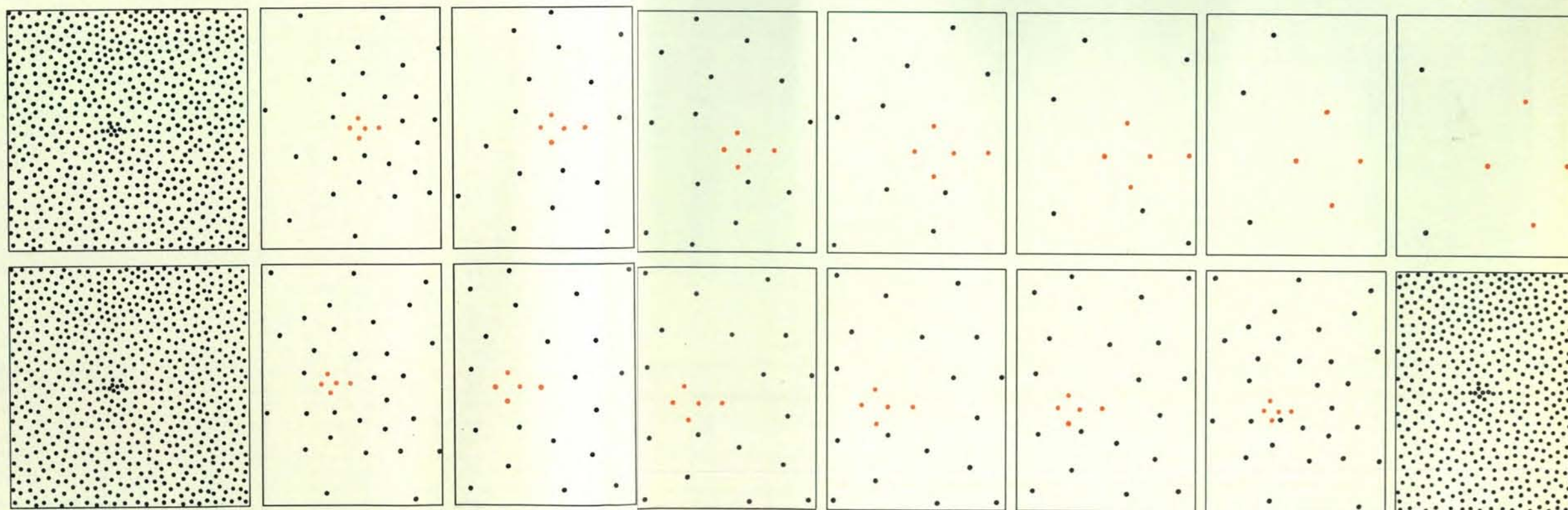
Il moto di recessione è dotato di varie proprietà interessanti. Hubble dimostrò che la velocità con cui una galassia si allontana è proporzionale alla sua distanza da noi, il che permette di determinare un rapporto costante tra la distanza e la velocità. Tale rapporto prevede che una galassia che si trova a 10 milioni di anni luce da noi si allontani alla velocità di 170 chilometri al secondo; un'altra galassia che disti il doppio si allontanerà con velocità doppia, cioè a 340 chilometri al secondo (*si vedano le illustrazioni delle due pagine seguenti*). Si osservano generalmente piccole deviazioni da questa norma, poiché la maggior parte delle galassie fa parte di gruppi o ammassi e pertanto presenta moti orbitali con una componente della velocità lungo la linea di vista che unisce la Terra con la galassia. Questi moti hanno però natura casuale, pertanto si cancellano reciprocamente in qualsiasi campione che contenga un gran numero di galassie. Deviazioni non casuali, sistematiche, dalla legge di proporzionalità sono state trovate solo per galassie estremamente distanti; come vedremo queste deviazioni non privano di validità la legge, ma forniscono importanti informazioni sulla storia dell'universo.

Una seconda caratteristica dell'espansione cosmica è la sua isotropia: essa infatti procede ugualmente in ogni direzione. La velocità di recessione di una galassia è correlata alla sua distanza mediante la stessa costante di proporzionalità indipendentemente dalla posizione occupata sulla volta celeste. Questa osservazione sembra indicare che l'universo è dotato di elevata simmetria e che, cosa ancor più straordinaria, noi ci troviamo proprio nel suo centro. Le cosmologie medievali, con l'ipotesi delle sfere cristalline, non erano maggiormente geocentriche.

Naturalmente c'è un'altra spiegazione, che può essere facilmente compresa considerando un semplice modello bidimensionale dell'universo in espansione. Immaginate un palloncino sferico sulla cui superficie sono dipinti piccoli punti, ciascuno dei quali rappresenta una galassia. Gonfiando il palloncino, la distanza tra due punti qualsiasi (misurata sempre sulla superficie della sfera) cresce con una velocità proporzionale alla distanza reciproca. Indipendentemente dal punto che si è voluto scegliere come centro, tutti gli altri punti si allontanano da esso uniformemente in tutte le direzioni. Pertanto da ogni punto si osserva un'espansione della stessa intensità e nessun punto ha una posizione privilegiata. Un'espansione così fatta non ha centro, o meglio, ogni punto è il suo centro.

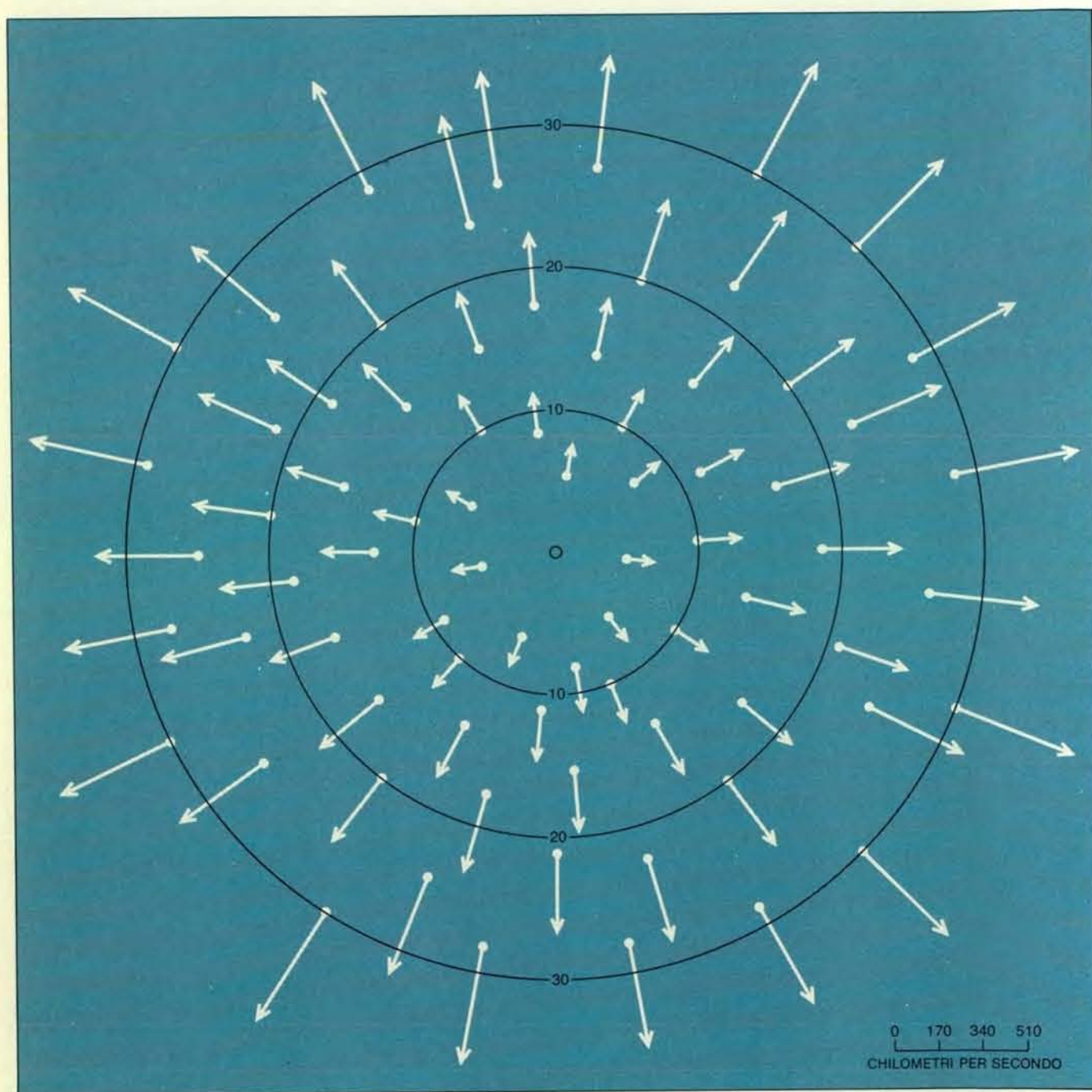
Da questa analisi dell'espansione si deduce che la configurazione geometrica dei punti non può cambiare. Un palloncino su cui è dipinta un'immagine di Topolino continua a mostrare la stessa figura quando è gonfiato. Tutte le distanze fra i punti sul palloncino sono cresciute di uno stesso fattore moltiplicativo. Analogamente, nell'universo reale, otto galassie che in un dato istante sono collocate sui vertici di un cubo, rimangono sui vertici di un cubo, seppure più grande, mentre l'universo si espande.

Il «big bang»

Dopo la scoperta iniziale di Hubble, osservazioni sempre più precise hanno mostrato che l'isotropia non è una proprietà esclusiva dell'espansione cosmica; tutte le caratteristiche dell'universo osservabili su larga scala sono indipendenti dalla direzione. Per esempio, la distribuzione delle galassie sulla sfera celeste e la distribuzione delle radiosorgenti extragalattiche appaiono del tutto uniformi. La prova decisiva a favore dell'isotropia fu fornita nel 1965 da Arno A. Penzias e Robert W. Wilson dei Bell Laboratories; è la radiazione di fondo nella banda delle microonde che sembra permeare tutto

Due classi di modelli dell'evoluzione dell'universo sono in genere ritenute verosimili; in entrambe l'universo ha inizio in uno stato compatto a densità infinita (il big bang). In una classe l'universo si espande continuamente e indefinitamente, sebbene a velocità sempre minore (*serie superiore di figure*). Nell'altra l'universo si espande fino a raggiungere un'estensione massima, poi comincia a contrarsi ricadendo alla fine in uno stato di densità infinita (*serie inferiore di figure*). I due casi sono illustrati qui per una regione arbitraria dello spazio in cui l'espansione è rappresentata da una riduzione della densità. L'espansione è isotropa, ovvero è la stessa in ogni direzione, e quindi un osservatore, indipendentemente dalla posizione occupata, percepisce se stesso al centro dell'espansione e lo schema di una configurazione geometrica (come quella arbitraria indicata nella figura coi punti colorati) rimarrà invariato in ogni epoca.

l'universo. Si è dimostrato che questa radiazione è fortemente isotropa: la sua variazione sull'intera area del cielo è minore dell'uno per mille.

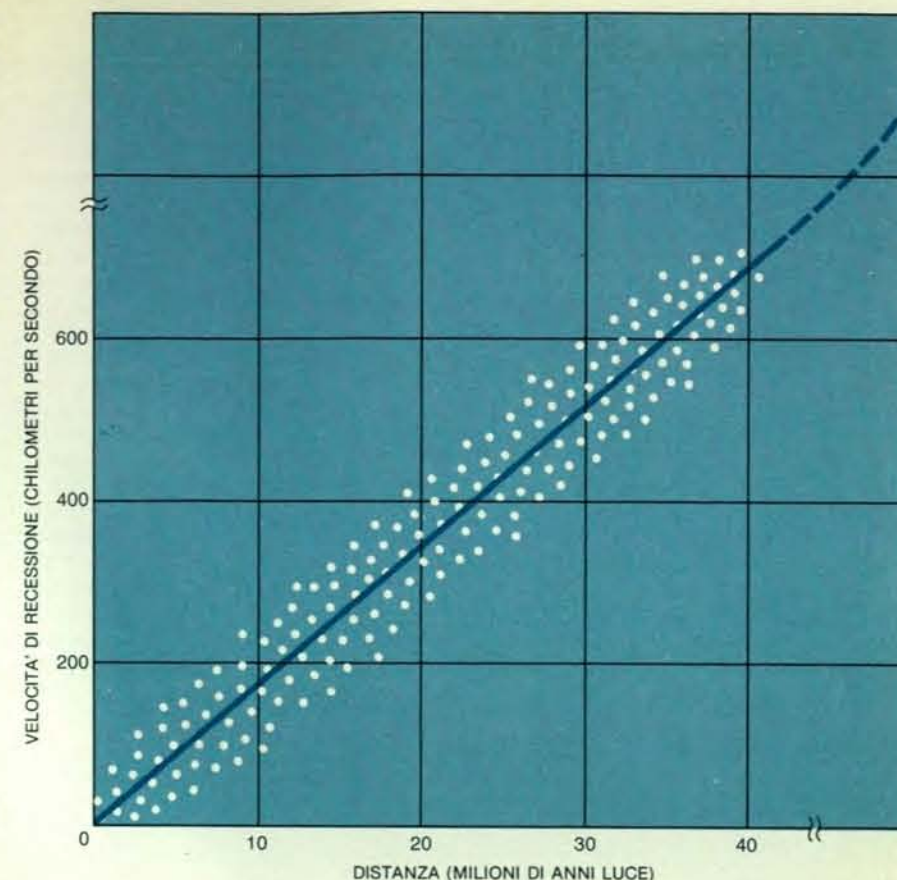
L'osservazione di un grado così elevato di isotropia ha portato alla formulazione di una legge generale di enorme importanza nota come il principio cosmologico, la quale afferma che in ogni luogo e in ogni epoca l'universo appare isotropo a qualsiasi osservatore che partecipi all'espansione. In altri termini, la nostra galassia è davvero nel centro dell'universo, ma condivide questo privilegio con ogni altra galassia.

Il principio cosmologico determina anche il comportamento del modello bidimensionale di universo rappresentato da un palloncino sferico. Se i punti dipinti sono distribuiti con densità uniforme sulla superficie del palloncino, allora l'aspetto della regione prossima a un punto fissato è statisticamente lo stesso indipendentemente dal punto considerato e non esistono direzioni privilegiate. In effetti non è necessario formulare indipendentemente l'ipotesi che i punti (ovvero, nell'universo tridimensionale, le galassie) siano distribuiti uniformemente. Infatti se l'universo è isotropo per ogni osservatore, allora la distribuzione deve essere omogenea; altrimenti un osservatore che si trovi ai confini di una fluttuazione di densità non vedrebbe una distribuzione uniforme indipendentemente dalla direzione.

A causa dell'isotropia, l'espansione cosmica sembra porre l'osservatore al centro dell'universo, nel punto cioè da cui fuggono via tutte le galassie lontane. La velocità di recessione di una galassia è proporzionale alla sua distanza dall'osservatore. Questa relazione è stata determinata per la prima volta negli anni venti da Edwin P. Hubble mediante osservazioni condotte col telescopio da 100 pollici del Mount Wilson Observatory. Da allora l'affermazione della costanza del rapporto tra la velocità di recessione e la distanza è nota come la legge di Hubble. Il modo più semplice per interpretare questo rapporto è pensare che l'espansione abbia avuto inizio con il big bang, dal momento che la relazione esistente implica che in passato tutte le galassie formavano un blocco molto compatto di densità infinita. Le distanze sono espresse in milioni di anni luce; le velocità sono rappresentate dalle lunghezze delle frecce misurate secondo la scala in basso a destra.

Per gli scopi della nostra trattazione adotteremo il principio cosmologico, ma si deve ricordare che il suo interesse è essenzialmente speculativo, dato che non è stato verificato in modo soddisfacente e, in effetti, può darsi che non sia suscettibile di prove conclusive.

Data la nostra conoscenza dell'universo basata sulle osservazioni odierne, che cosa possiamo dedurre circa la sua storia? Possiamo incominciare con un modello semplice che suppone che la velocità di recessione di ogni galassia sia rimasta invariata nel tempo. Da ciò si può derivare che la distanza che ci separa da qualsiasi galassia in recessione diminuisce gradualmente man mano che si procede verso il passato, e che il tempo trascorso dallo stato di massima prossimità è uguale al rapporto tra la distanza della galassia e la sua velocità. Ma poiché tale rapporto è lo stesso per tutte le galassie, allora tutte quante devono essere state vicine nello stesso tempo; in altre parole, in un particolare istante del passato tutta la materia dell'universo era compressa in una massa di altissima densità. Il tempo trascorso a partire da quello stato di massima compattezza, calcolato nell'ipotesi che la velocità di espansione non sia variata, è detto tempo di Hubble. Il suo reciproco, per il quale va moltiplicata la distanza di una galassia per ottenerne la velocità di recessione, è detto la costante di Hubble. La determinazione del tempo di Hubble è resa difficile da numerosi elementi di incertezza sulla distanza delle galassie, e i valori delle misurazioni sono stati corretti più volte dopo una prima stima pari a circa due miliardi di anni dovuta a Hubble. Oggi si ritiene che il tempo di Hubble sia compreso tra 12 e 25 miliardi di anni, e che il valore più probabile sia di circa 19 miliardi.

Se si estrapolano nel passato i moti delle galassie, si giunge alla fine a uno stato in cui tutte le galassie sono schiacciate l'una sull'altra con una densità infinita. Questa situazione rappresenta il *big bang* (grande esplosione), e segna l'origine dell'universo e di ogni cosa in esso. Con un semplice calcolo si vede che, se le velocità di recessione non sono cambiate, allora il giorno del big bang deve essere stato esattamente un tempo di Hubble fa. In realtà la velocità di espansione quasi certamente non è rimasta costante, ma questo non contraddice il fatto che ci sia stato il big bang; semplicemente ne sposta la data.

Il fatto che l'universo abbia avuto origine con una grande esplosione è una conclusione inevitabile se si suppone che le leggi note della fisica siano corrette e in certa misura complete. Non va trascurata però la possibilità che esistano leggi di natura i cui effetti non sono apprezzabili su scale dell'ordine di quelle dei laboratori di fisica e nemmeno di quelle del sistema solare, ma che possano predominare nel determinare il comportamento dell'universo nella sua totalità. Una teoria che fa ricorso a nuove leggi di natura è la cosmologia dello stato stazionario, in cui l'universo non muta il suo aspetto nel tempo e ha età infinita.

La legge di Hubble si determina misurando il rapporto tra la velocità e la distanza per molte galassie. La stima migliore di tale rapporto (*linea continua in colore*) è di circa 17 chilometri al secondo ogni milione di anni luce. Le singole galassie (*punti bianchi*) si discostano da tale valore poiché la maggior parte di loro fa parte di ammassi e presenta velocità orbitali. L'inverso del rapporto è il tempo di Hubble: cioè il tempo impiegato da ciascuna galassia per raggiungere la posizione attualmente occupata se si fosse mossa sempre con velocità pari a quella odierna, ovvero, il tempo che sarebbe trascorso dal big bang se le velocità non fossero cambiate. In realtà si ritiene che le velocità di recessione siano diminuite a causa dell'attrazione gravitazionale; si ritiene perciò che il valore del rapporto cresca considerando distanze grandissime (*linea tratteggiata in colore*).

Per rendere conto dell'espansione cosmica la teoria dello stato stazionario ipotizza la creazione di materia dal nulla.

Nel modello di universo dello stato stazionario è particolarmente arduo spiegare la radiazione di fondo nelle microonde. Questo campo di radiazione ha le caratteristiche spettrali della radiazione termica emessa da un corpo nero alla temperatura di 2,7 kelvin, e sembra spiegato in modo soddisfacente solo come residuo di un'epoca in cui l'universo era molto caldo e denso. Un universo a stato stazionario non può essere stato in una condizione siffatta, dal momento che in tale modello tutte le condizioni sono invariabili per definizione.

Nei modelli che prevedono il big bang la radiazione di fondo è una conseguenza naturale delle condizioni esistenti nell'universo primitivo. In questi modelli lo stato iniziale è caratterizzato da elevata temperatura e densità, uno stato detto talvolta «palla di fuoco cosmica». Si pensa che in tale stadio la materia e l'energia elettromagnetica che componevano l'universo fossero in equilibrio termodinamico: pertanto lo spettro della radiazione era quello di un corpo molto caldo. Man mano che l'universo si espandeva la radiazione si raffreddava, fino a giungere allo spettro a bassa temperatura osservato oggi. Tale raffreddamento può essere interpretato come un colossale spostamento verso il rosso; dal momento che tutte le galassie si allontanano costantemente dalla radiazione, lo spettro di questa è costantemente spostato verso lunghezze d'onda maggiori, associate con energie minori e temperature più basse. Nel 1946 George Gamow previde l'esistenza di una radiazione di fondo di tipo termico deducendola essenzialmente dall'apparato teorico del modello del big bang. Egli calcolò per la temperatura attuale della radiazione di fondo il valore di cinque kelvin. L'accordo tra le previsioni di Gamow e le osservazioni di Penzias e Wilson è la prova più significativa a favore del big bang.

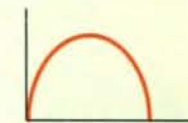
Si ritiene pertanto che l'universo sia nato da uno stato a densità infinita circa un tempo di Hubble fa. Lo spazio, il tempo e tutta la materia dell'universo furono creati allora. Non ha significato chiedersi che cosa ci sia stato prima del big bang: sarebbe un po' come domandarsi che cosa c'è a nord del polo nord. Analogamente non ha senso chiedersi dove abbia avuto luogo il big bang. L'uni-

verso puntiforme non era un oggetto isolato nello spazio, ma era l'universo intero, e perciò la sola risposta possibile a questa domanda è che il big bang è accaduto ovunque.

Nella maggior parte dei modelli dell'universo in evoluzione si suppone che le galassie in recessione seguano traiettorie balistiche, più o meno simili a quella di una palla lanciata o di un proiettile d'artiglieria. Le galassie sono state proiettate da forze che hanno agito nel momento del big bang, ma da allora in poi si sono mosse in volo libero, senza ulteriori spinte. Se su di esse non agissero altre forze dovrebbero perciò procedere con moto uniforme. In realtà le galassie continuano a interagire mentre si allontanano. Se nei nostri modelli sono accettabili solo le forze comuni con cui esprimiamo le leggi note della fisica, allora una sola forza può avere un effetto significativo sull'espansione: la gravitazione. Possiamo pertanto sperare di capire la dinamica di un universo in espansione se riusciamo a descrivere l'interazione gravitazionale di tutte le sue componenti.

Decelerazione gravitazionale

La forza gravitazionale agisce su tutta la materia, è sempre attrattiva e il suo raggio d'azione è infinito. Inoltre, la gravitazione gode di una particolare proprietà geometrica che ne semplifica notevolmente l'analisi: una sfera cava non esercita nessuna forza gravitazionale netta sulle masse poste nel suo interno. (In realtà, naturalmente, la massa del guscio attrae le masse che sono poste nel suo interno, ma le varie forze si annullano a vicenda, così che in ogni punto interno la forza risultante è nulla.) Questa affermazione fu provata per la prima volta da Newton, ma si applica altrettanto bene a teorie più moderne della gravitazione, come la teoria generale della relatività.

Se si prende in esame una regione sferica dell'universo, il resto dell'universo che la circonda può essere considerato come un guscio sferico, dal momento che il principio cosmologico implica che la materia circostante sia distribuita uniformemente in tutte le direzioni. La sfera in esame può allora essere studiata come se fosse isolata e non soggetta a forze esterne. Il principio cosmologico garantisce inoltre che qualsiasi sfera di galassie considerata si espanderà o contrarrà con la stessa legge di proporzionalità che si applica all'universo intero, indipendentemente dalla localizzazione e dalle dimensioni della sfera stessa. A seguito di ciò, per determinare la dinamica dell'universo, basterà analizzare la dinamica di una qualsiasi sfera campione in esso. Se la sfera scelta sarà piccola, le velocità delle galassie saranno molto minori della velocità della luce, e i loro moti potranno essere trattati coi metodi della meccanica newtoniana.

Una galassia che si trovi sul bordo della piccola regione sferica è soggetta solo alle forze gravitazionali generate dalla materia interna alla sfera. Se la distribuzione della materia è omogenea, allora la forza risultante che agisce sulla galassia la attrae verso il centro della sfera. A causa di ciò la galassia in esame non si allontana a velocità costante; il suo moto di recessione risulta invece ritardato. È ovvio pertanto che nel passato la galassia in esame, e tutte le altre galassie, si muovessero più velocemente di oggi. Se si trascura la decelerazione si sovrastima l'età dell'universo. Quest'ultima è pari a un tempo di Hubble solo nell'ipotesi che la velocità d'espansione non sia mutata; dal momento però che l'espansione è stata rallentata dall'azione della gravitazione, il big bang deve essere avvenuto in epoche più recenti di un tempo di Hubble fa.

L'importanza della decelerazione gravitazionale dipende ovviamente dalla quantità di materia che si trova all'interno della sfera considerata. Se la sfera contiene una grande quantità di materia, la galassia in esame alla fine sarà arrestata e cadrà verso il centro; la regione sferica campione comincerà a contrarsi e con essa, in forza del principio cosmologico, l'intero universo. Se c'è poca materia, la galassia in esame continuerà a rallentare, ma non si arresterà mai. Sia la regione sferica sia tutto l'universo si espanderanno indefinitamente. La situazione è simile a quella di un proiettile scagliato verso l'alto della superficie della Terra: il proiettile rallenta, ma ciononostante non ricadrà sulla superficie se la sua velocità è superiore a un certo valore critico, la velocità di fuga.

La velocità di fuga per oggetti che lasciano la Terra è determinata dalla massa della Terra stessa; per una galassia campione sulla superficie di una sfera arbitraria nello spazio la velocità di fuga è determinata dalla massa totale all'interno della sfera. La velocità di recessione odierna della galassia campione rispetto al centro della sfera è nota dal rapporto velocità-distanza. Il destino della galassia dipende perciò dal valore della velocità di fuga e quindi dalla massa interna alla sfera.

Dal momento che si fa l'ipotesi che l'universo sia omogeneo, la quantità da determinarsi è la densità media della materia nell'universo. Se la densità è minore di un certo valore critico, l'effetto

I modelli dell'evoluzione cosmica descrivono la variazione delle dimensioni dell'universo al passare del tempo. Tutti i modelli devono essere compatibili con le dimensioni e la velocità di espansione osservate oggi, così che tutte le loro rappresentazioni grafiche devono essere tangenti tra loro all'istante attuale. Se la velocità di espansione è costante (*linea nera tratteggiata*), l'età dell'universo è uguale al tempo di Hubble. Tutti gli universi che prevedono una decelerazione sono più giovani, e sia la loro storia passata sia il loro destino futuro dipendono dal valore della decelerazione. Se il rallentamento è relativamente piccolo, l'espansione può continuare indefinitamente, anche se con velocità sempre minore (*linea nera continua*). Un rallentamento maggiore implica che l'espansione cosmica deve arrestarsi e poi invertirsi, portando a un collasso finale (*linea in colore*). L'universo che si espande indefinitamente è detto «aperto»; l'universo destinato a collassare, che è anche il più giovane, tra quelli che abbiamo rappresentato, è detto «chiuso».

		APERTO	CRITICO	CHIUSO
PARAMETRO DI DENSITA' Ω	$\frac{\text{DENSITA' EFFETTIVA}}{\text{DENSITA' CRITICA}}$	$\Omega < 1$	$\Omega = 1$	$\Omega > 1$
PARAMETRO DI DECELERAZIONE q_0	DECELERAZIONE $\frac{\text{DISTANZA}}{(\text{VELOCITA'})^2}$	$q_0 < \frac{1}{2}$	$q_0 = \frac{1}{2}$	$q_0 > \frac{1}{2}$
	GEOMETRIA DELLO SPAZIO	IPERBOLICA (CURVATURA NEGATIVA)	PIATTA (CURVATURA NULLA)	SFERICA (CURVATURA POSITIVA)
	FUTURO DELL'UNIVERSO	ESPANSIONE PERPETUA	ESPANSIONE PERPETUA	COLLASSO FINALE

I modelli aperti e chiusi dell'universo sono distinti soprattutto dalla densità media della materia e dal valore della decelerazione cosmica. La densità è un fattore decisivo poiché nei modelli descritti dalla teoria generale della relatività essa sola fissa il valore delle forze gravitazionali che rallentano l'espansione cosmica. La densità è descritta nel modo più semplice in termini di un parametro adimensionale: il rapporto tra la densità effettiva e la densità critica, dove quest'ultima corrisponde al minimo valore necessario per arrestare l'espansione. Anche la decelerazione può essere espressa mediante una grandezza adimensionale, il parametro di decelerazione, che nei modelli qui considerati è sempre uguale a metà del parametro di densità. Questi due parametri determinano non solo il futuro dell'universo, ma anche la geometria dello spazio. L'universo aperto ha estensione infinita in ogni istante e il suo spazio ha curvatura iperbolica negativa. Nell'universo che ha densità uguale a quella critica, nel quale il parametro di densità è esattamente uguale a 1, lo spazio ha curvatura nulla: è lo spazio piatto della geometria euclidea. L'universo chiuso ha estensione finita; in tale universo lo spazio ha curvatura di tipo sferico positiva.

della gravitazione è troppo debole per arrestare l'espansione cosmica, e tutte le galassie continueranno sempre ad allontanarsi (seppure sempre più lentamente). Se la densità è maggiore della densità critica, allora la gravitazione avrà la meglio: l'espansione sarà rallentata fino a fermarsi, comincerà poi una contrazione accelerata che porterà a una catastrofe finale, che potrebbe essere detta il *big crunch* (grande implosione). Il valore effettivo della densità critica è funzione del tempo di Hubble, che non è noto con precisione. Se il tempo di Hubble è di 19 miliardi di anni la densità critica è di 5×10^{-30} grammi per centimetro cubo, equivalente a circa tre atomi di idrogeno per metro cubo. Sembra una densità straordinariamente piccola, ma si deve tenere presente che in media l'universo è praticamente vuoto.

Il modo più conveniente per inserire nei modelli matematici l'effetto della gravitazione sull'espansione cosmica consiste nell'introdurre una grandezza adimensionale detta parametro di densità e indicata con la lettera greca omega (Ω). Il parametro di densità è definito come il rapporto tra la densità effettiva dell'universo e la densità critica. Perché l'universo continui a espandersi per sempre bisogna che tale rapporto sia minore o uguale a uno; se Ω è esattamente uguale a uno, l'universo si sta espandendo ovunque con velocità esattamente uguale a quella di fuga; se Ω è maggiore di 1, l'universo dovrà alla fine collassare.

La geometria dello spazio

La trattazione svolta fin qui poteva essere dedotta completamente dalla teoria newtoniana della gravitazione, nonostante sia valida anche nella teoria generale della relatività. Nella teoria generale, però, il valore del parametro di densità determina ulteriori conseguenze; in particolare fissa la geometria dello spazio. Nell'universo ad alta densità destinato al collasso la gravitazione è sufficientemente intensa da «chiudere» lo spazio. Il volume totale dell'universo è finito in ogni istante, sebbene l'universo non abbia né confini né bordi. Un analogo bidimensionale di uno spazio tridimensionale siffatto è la superficie di una sfera, che pure ha area finita, ma non ha confini.

Se Ω è uguale a 1 così che l'universo si espande proprio con la velocità di fuga, la geometria dello spazio è «piatta», ossia la comune geometria euclidea, ed è rappresentata in due dimensioni da un piano infinito.

La geometria di un universo in espansione perpetua, in cui Ω è minore di 1, è più difficile da descrivere. Il corrispondente bidimensionale è la superficie di una figura detta pseudosfera, ma una pseudosfera completa non può essere costruita in uno spazio a tre dimensioni. Una superficie a forma di sella gode di alcune delle proprietà di tale spazio. L'analogia è però imperfetta in un aspetto molto importante: la superficie a sella ha un centro, mentre lo spazio reale non ammette posizioni privilegiate (*si veda l'illustrazione della pagina seguente*). Forse la migliore rappresentazione bidimensionale di uno spazio di tale genere è la proiezione di una pseudosfera su un piano, una figura che è stata utilizzata in varie opere dell'artista olandese M.C. Escher (*si veda l'illustrazione a pagina 49*).

Le tre specie possibili di spazio tridimensionale sono distinte da varie proprietà geometriche, alcune delle quali possono essere rappresentate nei modelli bidimensionali. Un piano privo di curvatura è ovviamente la base della geometria di Euclide e per esso valgono tutti gli assiomi euclidei e i teoremi che si deducono da essi. Su un piano può essere tracciata una e una sola retta che passi per un punto assegnato e sia parallela a

Alla geometria dello spazio caratteristica di ciascun modello di universo si può fare corrispondere una superficie bidimensionale analoga. Le proprietà delle superfici sono caratterizzate dagli assiomi e dai teoremi di Euclide sulle linee parallele, sulla somma degli angoli interni di un triangolo e sulla misura della lunghezza della circonferenza e dell'area di un cerchio. Lo spazio piatto di un universo critico è rappresentato da un piano, e lo spazio a curvatura positiva di un universo chiuso corrisponde alla superficie di una sfera. Alcune delle proprietà dello spazio a curvatura negativa di un universo aperto possono essere dimostrate per una superficie a forma di sella, ma questa analogia è imperfetta, perché la sella ha un centro. La migliore rappresentazione di un universo aperto è una superficie infinita detta pseudosfera, che però non può essere costruita in uno spazio a tre dimensioni.

una retta data; la somma degli angoli interni in un triangolo è sempre 180 gradi; la circonferenza di un cerchio è direttamente proporzionale al raggio e l'area di un cerchio è direttamente proporzionale al quadrato del raggio.

Sulla superficie di una sfera non esistono due rette parallele tra loro, quando per linea retta si intenda quella che collega due punti seguendo il cammino più breve. Linee così definite sono dette geodetiche. Sulla sfera le geodetiche sono cerchi massimi e due qualsiasi di loro si intersecano sempre. Inoltre su una sfera la somma degli angoli interni in un triangolo è sempre maggiore di 180 gradi; la circonferenza di un cerchio cresce più lentamente del raggio e l'area di un cerchio cresce più lentamente del quadrato del raggio.

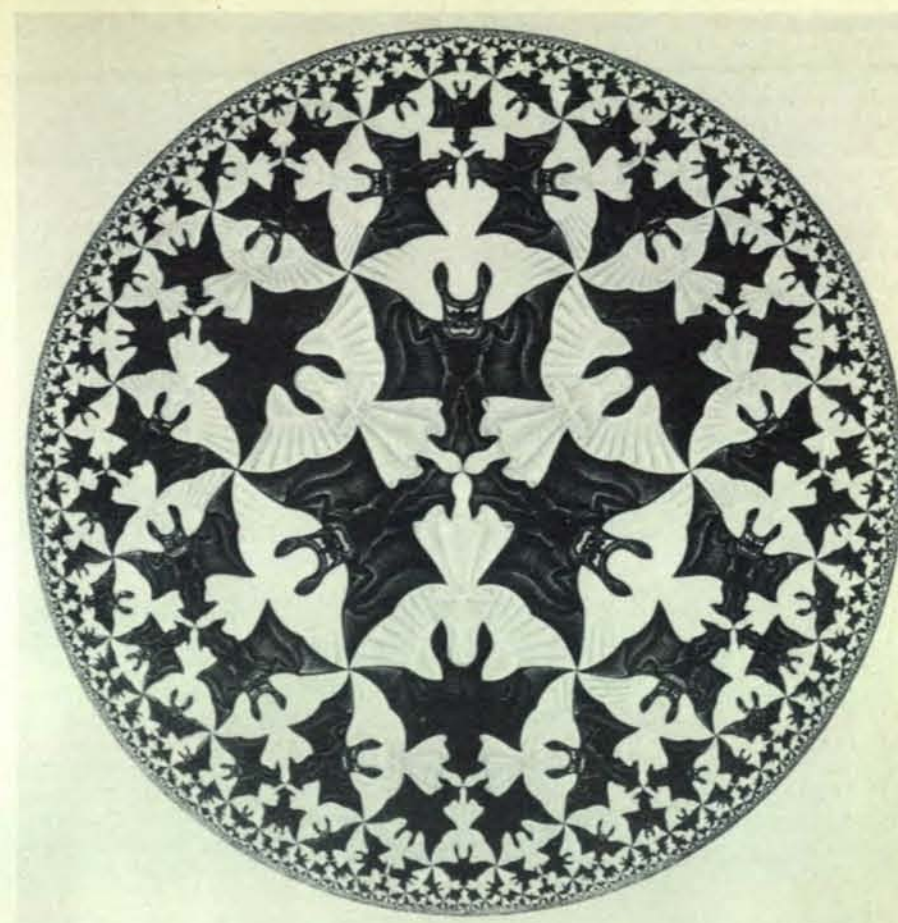
La superficie di una pseudosfera gode di proprietà opposte a quelle della sfera. Per un punto assegnato possono essere tracciate infinite rette parallele a un'altra retta, o geodetica. La somma degli angoli interni di un triangolo è minore di 180 gradi. La circonferenza di un cerchio cresce più velocemente del raggio e l'area di un cerchio cresce più velocemente del quadrato del raggio. La geometria dello spazio tridimensionale rappresentato da una pseudosfera fu studiata per primo da Nikolai Lobachevski nel 1826.

Nei semplici modelli cosmologici che trattiamo qui la geometria dello spazio è connessa in modo univoco al comportamento futuro dell'universo. È interessante osservare che nei modelli con Ω maggiore di 1 l'universo è chiuso sia spazialmente sia temporalmente. Il volume dello spazio è finito e ci sono limiti temporali definiti: dal big bang iniziale al *big crunch* finale. I modelli in cui Ω è minore o uguale a 1 sono aperti sia nello spazio sia nel tempo. Tali modelli hanno un punto di inizio definito (il big bang), ma sono sempre infinitamente estesi e si sviluppano nel futuro.

Misure di decelerazione

Esistono vari metodi per cercare di determinare se l'universo fisico è aperto o chiuso. Tutti portano alla fine a una stima del tasso di rallentamento dell'espansione cosmica. Un metodo consiste semplicemente nel misurare la decelerazione direttamente, osservando galassie lontane. È pure possibile misurare l'età dell'universo, e da quella, per confronto col tempo di Hubble (età in assenza di decelerazione), stimare di quanto è variata la velocità di espansione. Dal momento che la decelerazione è un fenomeno gravitazionale, una misura equivalente è quella della densità media della materia; confrontando la densità effettiva con la densità critica si ottiene il rapporto Ω. Infine, l'abbondanza odierna di certi elementi chimici rappresenta una sorta di registrazione fossile delle condizioni, densità compresa, esistenti nelle primissime fasi dello sviluppo dell'universo; anche da questa informazione è possibile calcolare il valore di Ω. Le indicazioni che è possibile trarre da tutti questi metodi hanno contribuito a formare la nostra conoscenza attuale dello stato dell'universo.

Solitamente la decelerazione dell'espansione cosmica è espressa in termini di una quantità adimensionale detta parametro di decelerazione indicato con q_0. Dato che il rallentamento è un effetto gravitazionale il parametro di decelerazione è strettamente connesso con la densità media della materia. Nei modelli cosmologici qui considerati, che sono costruiti sulla base della teoria generale della relatività, q_0 è sempre esattamente uguale alla metà del parametro di densità Ω. Pertanto se q_0 è maggiore di 1/2, allora l'universo, a causa della sua elevata densità, decelera abbastanza rapidamente da smettere di espandersi, per poi collassare. Se q_0 è minore di 1/2 l'espansione non può avere termine, perché la densità è troppo bassa per arrestarla.

Un metodo ovvio per determinare il tasso di decelerazione consisterebbe nel misurare la componente radiale della velocità di una data galassia in due tempi diversi, così da stabilire di quanto è stata rallentata nell'intervallo tra le due misure. Purtroppo però la variazione di velocità prevista in periodi di durata comparabile a quella della vita umana è troppo piccola per essere misurata; infatti gli errori sperimentali connessi a tale misura superano di vari ordini di grandezza il valore previsto. D'altra parte, il fatto che la velocità della luce è finita permette di misurare velocità di galassie in un lontano passato e di confrontarle poi con velocità corrispondenti a epoche più recenti. Il confronto è realizzabile perché guardando oggetti sempre più distanti nel cielo ci spingiamo anche sempre più indietro nel passato. La relazione è ovvia quando le distanze sono misurate in anni luce: se una galassia dista un miliardo di anni luce, allora la luce che noi riceviamo oggi da essa è stata emessa un miliardo di anni fa, e lo spostamento Doppler nel suo spettro deve riflettere la velocità che la galassia aveva allora relativamente alla nostra velocità odierna. Pertanto, se l'espansione cosmica sta rallentando, ci si aspetta che la costanza del rapporto tra velocità e distanza scoperta da Hubble non valga per le galassie più lontane.

La superficie di una pseudosfera è rappresentata in una xilografia, *Limite del cerchio IV*, di M.C. Escher. Nella xilografia la superficie è proiettata su un piano. Come accade nelle proiezioni cartografiche, la scala non è costante; sulla pseudosfera effettiva tutte le figure degli angeli e dei demoni avrebbero le stesse dimensioni. Se si prende una singola figura come unità di misura, appare chiaro che la circonferenza di un cerchio cresce molto più rapidamente del raggio. Analogamente, ogni figura definisce un triangolo (coi vertici in corrispondenza dei piedi e delle estremità delle ali): dal numero di triangoli che si incontrano in ogni vertice si può dimostrare che sulla pseudosfera la somma degli angoli interni di un triangolo è minore di 180 gradi. La pseudosfera è una superficie infinita a curvatura negativa, analoga allo spazio in un universo che si espande per sempre. In essa non c'è nessuna posizione privilegiata che possa essere presa come centro, e la proiezione rimarrebbe invariata se fosse centrata su qualsiasi altro punto.

Alle distanze più remote tale rapporto dovrebbe aumentare, cioè, in altre parole, le velocità osservate dovrebbero essere maggiori di quelle che sono previste dalla legge di Hubble.

Per misurare la decelerazione con questo metodo è necessario avere una misura indipendente delle distanze delle galassie. Escludendo le galassie più vicine, l'unico metodo utilizzabile per valutare la distanza di una galassia si basa sulla sua luminosità apparente. Se tutte le galassie in tutti i tempi avessero la stessa luminosità intrinseca, allora la loro brillanza apparente sarebbe semplicemente proporzionale all'inverso del quadrato della loro distanza, e la determinazione di quest'ultima sarebbe immediata. Naturalmente, esse non hanno tutte la stessa luminosità intrinseca.

Variazioni casuali della brillanza (dovute, per esempio, a diversità di dimensioni) possono dar luogo a errori nelle singole misure. A causa di tali variazioni bisogna accumulare una gran mole di dati e sottoporla ad analisi statistica; comunque, in linea di principio, le variazioni di natura casuale non costituiscono un problema serio, perché ci si attende che si cancellino reciprocamente se si dispone di un campionamento abbastanza vasto. Invece, le variazioni di natura sistematica richiedono una correzione esplicita per annullarne gli effetti.

Le teorie sull'evoluzione stellare indicano che la luminosità combinata di tutte le stelle in una galassia isolata diminuisce probabilmente di pochi centesimi in un miliardo di anni. Pertanto in un lontano passato le galassie erano probabilmente più brillanti. Se questa variazione di brillanza fosse trascurata nell'eseguire misure della decelerazione, le distanze calcolate sarebbero troppo piccole e, a causa di ciò, il tasso di decelerazione sarebbe sovrastimato. La riduzione della brillanza potrebbe sembrare molto modesta, ma essa cambia il valore calcolato del parametro di decelerazione q_0 di circa 1, cioè di una quantità più che sufficiente per decidere tra un universo aperto e uno chiuso. Le migliori osservazioni

di cui si dispone al momento attuale, nelle quali si tiene conto anche delle variazioni di luminosità intervenute nel tempo a causa dell'evoluzione stellare, suggeriscono concordemente che il valore di q_0 è più vicino a zero che a 1/2 e perciò che l'universo è aperto e in espansione perpetua.

Un'altra incertezza di notevole importanza pesa sulla determinazione della decelerazione. La maggior parte delle galassie osservate si trova raggruppata in ammassi relativamente densi, così che bisognerebbe tenere conto di possibili interazioni tra le galassie. Per esempio, è stato dimostrato recentemente che negli ammassi le galassie più grandi «inghiottono» le più piccole, con una conseguente variazione di luminosità e dimensioni. Non è ancora possibile valutare l'importanza di tale cambiamento, né sapere con sicurezza se, a causa di ciò, la luminosità misurata aumenta o diminuisce. Aggiungere stelle a una galassia dovrebbe renderla più brillante, ma in osservazioni cosmologiche si misura solo la luminosità della parte centrale della galassia. Se la galassia «cannibale» si dilata significativamente il numero di stelle che verrebbero a occupare la regione centrale potrebbe risultare ridotto. Ne risulterebbe quindi che la luminosità della galassia apparirebbe più debole.

L'età dell'universo

A causa delle incertezze statistiche e dell'incompletezza della nostra conoscenza dell'evoluzione delle galassie, il valore di q_0 dedotto dalle misure della velocità di allontanamento è molto incerto. Basandosi solo su questo fattore non si può concludere che q_0 è minore di 1/2 e l'universo è aperto; sembra comunque che valori molto grandi di q_0, come q_0 uguale a 2, possano essere esclusi.

Il secondo modo per determinare il destino dell'universo consiste nel misurarne l'età. Se l'espansione non fosse rallentata, l'età sarebbe uguale al tempo di Hubble. Dato però che è rallentata, l'universo deve avere un'età inferiore al tempo di Hubble. Trovando la differenza tra l'età effettiva e il tempo di Hubble è possibile in teoria calcolare il parametro di decelerazione di q_0.

L'età dell'universo può essere valutata in due modi; entrambi forniscono solo limiti inferiori, poiché misurano le età di oggetti nell'universo, sebbene tali oggetti siano stati formati probabilmente entro il primo miliardo di anni dopo il big bang. Il primo metodo consiste nel determinare l'età delle stelle più vecchie tuttora osservabili. Si pensa che le stelle più vecchie abbastanza vicine da permettere osservazioni dettagliate siano quelle che si trovano negli ammassi globulari associati alla nostra galassia. I modelli di evoluzione stellare indicano che hanno età comprese tra otto e 16 miliardi di anni.

L'età può essere stimata anche da misure dell'abbondanza relativa di certi elementi pesanti. Si ritiene che tutti gli elementi più pesanti del ferro, tra cui numerosi elementi radioattivi, siano stati sintetizzati nelle supernove, che probabilmente esplodono nella Galassia fin dalla sua formazione. Dato che ciascun elemento radioattivo decade con un tasso costante, il rapporto tra l'abbondanza di ogni elemento radioattivo e quella dei suoi prodotti di decadimento può dare indicazioni sull'età media degli elementi pesanti. I valori di questi rapporti suggeriscono che l'età della Galassia è compresa tra sei e venti miliardi di anni. Le due età calcolate sono dunque compatibili e indicano che il big bang ha avuto luogo in un periodo compreso tra otto e diciotto miliardi di anni fa.

Densità media

Il fatto che una data età, compresa nell'intervallo permesso, corrisponda a un universo aperto o a uno chiuso dipende dal valore del tempo di Hubble, che, come abbiamo visto, non è di facile determinazione. Inoltre, anche se si assume che il tempo di Hubble sia uguale alla migliore stima odierna di 19 miliardi di anni, né i limiti posti sull'età dell'universo, né l'esclusione dei valori di q_0 maggiori di 2 bastano per decidere se l'universo è aperto o chiuso (*si veda l'illustrazione nella pagina a fronte*). La questione può essere risolta solo imponendo ulteriori limitazioni.

Il terzo metodo consiste nel misurare la densità media dell'universo per ricavare da essa il parametro di densità Ω. È possibile ottenere un limite inferiore per la densità prendendo in considerazione solo la massa che è associata alle galassie visibili. La densità si ricava contando le galassie che si trovano in un dato volume di spazio, moltiplicando per la massa delle galassie e dividendo il risultato per il volume.

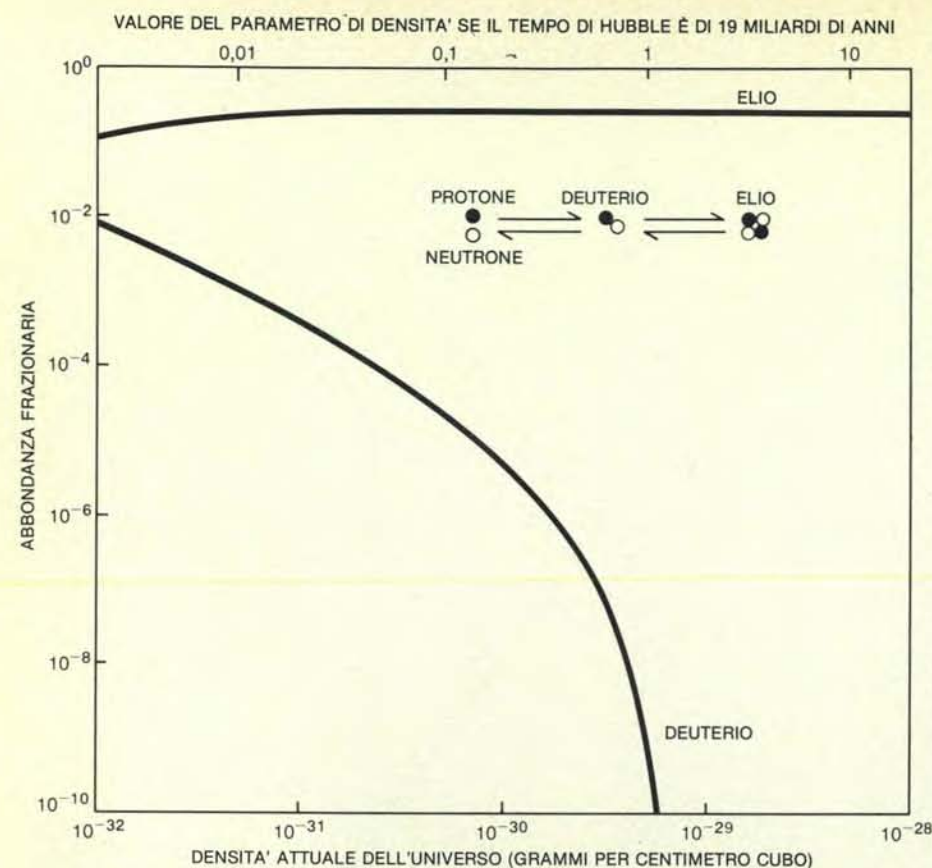
La valutazione della massa contenuta in una galassia non è così difficile come potrebbe sembrare a prima vista. Poche galassie sono completamente isolate; la maggior parte è riunita in piccoli gruppi o in grandi ammassi, e le loro masse possono essere dedotte osservando gli effetti della reciproca interazione gravitazionale. Per esempio, due galassie in orbita l'una intorno all'altra devono essere soggette a un'interazione gravitazionale capace di bilanciare la forza centrifuga. Se sono note la distanza reciproca e le loro velocità relative, la determinazione della loro massa combinata si riduce a un semplice esercizio di meccanica newtoniana. Il procedimento per ammassi composti da molte galassie è leggermente più complesso. La complicazione nasce essenzialmente dal fatto che la massa calcolata secondo questo metodo comprende non solo la massa della materia

La decelerazione dell'espansione cosmica può essere rivelata dalla conoscenza delle velocità di recessione delle galassie in un lontano passato. È possibile guardare nel passato osservando le galassie più distanti, infatti la luce che ci giunge oggi è stata emessa un numero di anni fa pari alla distanza della galassia misurata in anni luce. La decelerazione è percepita come una deviazione dalla legge di Hubble; infatti, se non ci fosse decelerazione il rapporto tra velocità e distanza sarebbe costante (*linea in nero*); in presenza di decelerazione il rapporto cresce alle distanze più grandi (*linea in colore*). A causa delle difficoltà che sono connesse alla valutazione delle distanze delle galassie non è stato possibile misurare con precisione il tasso di decelerazione, sono stati tuttavia esclusi valori del parametro di decelerazione maggiori o uguali circa a 2.

Alcune limitazioni sullo stato dell'universo sono fornite dalla determinazione della sua età e del parametro di decelerazione. Stime dell'età delle stelle più vecchie e dell'età media degli elementi pesanti indicano che l'universo ha da otto a 18 miliardi di anni; le corrispondenti limitazioni poste sul tempo di Hubble dipendono dalla densità. Le osservazioni delle galassie lontane permettono di porre un limite superiore al parametro di decelerazione; esso non può essere maggiore di 2 e quindi il parametro di densità non può essere maggiore di 4. Le limitazioni dedotte da queste sole misure non permettono di stabilire se l'universo è aperto o chiuso essendo compatibili coi due tipi di modelli.

che costituisce le galassie, ma anche la massa della materia che è presente nell'ammasso sotto ogni altra forma. In tal modo si tiene conto automaticamente anche di costituenti dell'universo che non sarebbero visibili, come buchi neri o polvere e gas extragalattici.

Le stime della massa di moltissime galassie, combinate coi conteggi delle galassie presenti in grandi volumi di spazio, danno un'indicazione del valore del parametro di densità Ω. Se la massa associata alle galassie rappresentasse tutta la massa dell'universo, allora Ω varrebbe solo 0,04 e l'universo sarebbe aperto e destinato a espandersi per sempre. Su questo valore pesa un'incertezza pari a circa un fattore 3, così che valori di Ω fino a 0,12 sono ancora compatibili con le osservazioni; ma anche così siamo ben al di sotto del valore di 1 necessario per chiudere l'universo.

La densità dell'universo può essere stimata anche confrontando il comportamento delle galassie lontane con quello delle galassie che fanno parte del super-ammasso locale, cioè del sistema di galassie che comprende il nostro gruppo locale insieme con molti altri piccoli raggruppamenti e con l'ammasso della Vergine, che ha dimensioni un po' più grandi. All'interno del super-ammasso locale la densità media delle galassie supera di circa due volte e mezza quella nell'universo totale. Se tutta la massa è associata alle galassie, allora anche la densità media della materia deve essere due volte e mezza più grande nel super-ammasso che al di fuori di esso. La differenza di densità dovrebbe produrre una differenza nel tasso di espansione; infatti, dato che la densità locale è maggiore, le galassie vicine dovrebbero essere maggiormente rallentate. L'effettiva importanza della diversità di decelerazione dipende dal valore di Ω; se Ω è grande ci sarà una differenza significativa tra le decelerazioni entro e fuori del super-ammasso locale. Se Ω è piccolo, allora il rallentamento è piccolo ovunque, e anche un incremento locale della densità pari a un fattore 2,5

La densità nelle prime fasi dell'universo influì sulla sintesi del deuterio e dell'elio, e dall'abbondanza relativa di quegli elementi si può ricavare la densità attuale. Si pensa che il deuterio si sia formato dalla fusione di protoni e neutroni nei minuti successivi al big bang, ma, se la densità fosse stata troppo grande, la maggior parte o tutto il deuterio sarebbe stato trasformato in elio. L'abbondanza di entrambi gli elementi è rappresentata come frazione (in termini di massa) di tutta la materia nell'universo. Se i modelli più semplici dell'universo primordiale sono corretti, e se il deuterio non è stato sintetizzato in eventi più recenti, l'abbondanza osservata indica come la densità dell'universo non possa superare il valore approssimato di 4×10^{-31} g/cm^3.

produrrà solo una piccola variazione. In effetti, la differenza non è rivelabile, essendo più piccola degli errori sperimentali probabili. La conclusione che si può trarre immediatamente è che Ω ha un valore molto piccolo, verosimilmente non più grande di 0,1.

Entrambi i metodi per determinare la densità sono chiaramente limitati alla materia associata alle galassie e si prestano alla naturale obiezione che possano esistere nell'universo quantità non trascurabili di materia sotto altra forma. Questa eventualità non può essere esclusa, ma non ci sono nemmeno prove a suo favore.

Le teorie correnti indicano che gli ammassi di galassie si sarebbero formati in un universo in cui la materia era distribuita molto più regolarmente di oggi. I residui lasciati dalla formazione delle galassie sarebbero stati inglobati anch'essi negli ammassi. Pertanto tutte le particelle che non sono negli ammassi devono essersi trovate in condizioni molto particolari, dovevano cioè possedere quelle speciali e insolite proprietà di posizione e velocità iniziali che permettessero loro di sfuggire all'attrazione degli ammassi. Anche se una grande quantità di materia fosse ancor oggi distribuita uniformemente fuori dagli ammassi, dovrebbe cadere entro a essi in pochi miliardi di anni.

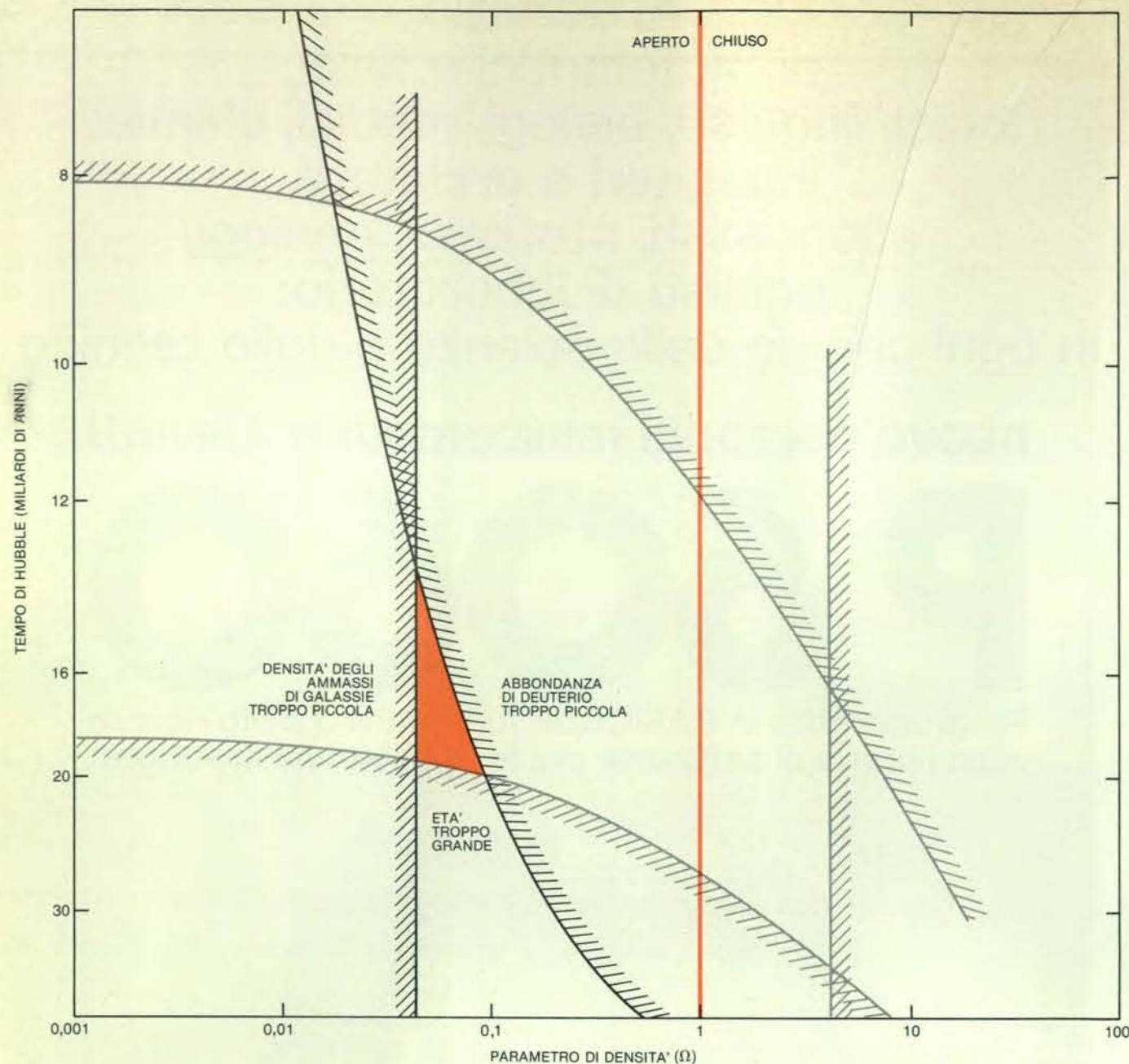
Alternativamente si potrebbe supporre che la massa necessaria per chiudere l'universo risieda in qualche mezzo uniformemente distribuito con una pressione interna sufficientemente elevata da renderlo praticamente insensibile all'attrazione gravitazionale delle galassie. Un mezzo con tali proprietà potrebbe essere composto, per esempio, da una grande quantità di neutrini o di onde gravitazionali. Si può però opporre una grave obiezione all'esistenza di questo mezzo di tipo radiativo che pervaderebbe l'universo: quasi certamente la sua presenza non avrebbe mai permesso che si formassero galassie e ammassi di galassie quali sono mostrati dalle attuali osservazioni.

In teoria si può determinare la densità di tutta la materia nell'universo, indipendentemente dal fatto che sia o meno associata con le galassie, ma solo estrapolando dalle condizioni nell'universo attuale a quelle esistenti pochi minuti dopo il big bang. Le ipotesi più semplici che si possono avanzare su quell'antichissimo periodo suggeriscono che la temperatura e la densità dovevano essere abbastanza elevate da permettere a qualche particella subatomica di interagire e formare quantità apprezzabili di alcuni dei nuclei più leggeri. In particolare un protone e un neutrone potrebbero fondersi in un nucleo di deuterio, e la maggior parte dei nuclei di deuterio si combinerebbe per formare nuclei di elio, composti da due protoni e due neutroni. La proporzione tra il deuterio e l'elio così sintetizzati dipende dalla densità dell'universo nel periodo in cui esso era abbastanza caldo da permettere la realizzazione di quelle reazioni. Dalla densità primitiva e dalla temperatura attuale della radiazione di fondo nelle microonde è possibile dedurre la densità odierna.

Densità primordiale

Modelli matematici indicano che, variando la densità dell'universo primitivo su tutto l'intervallo dei valori accettabili, si ottiene che la quantità di materia convertita in elio è compresa tra il 20 e il 30 per cento. L'abbondanza di elio misurata in tutta una serie di oggetti astronomici conferma la validità di questo stretto intervallo, rendendo quindi plausibile l'ipotesi fondamentale che l'universo abbia attraversato un periodo in cui densità e temperatura erano estremamente elevate poco dopo il big bang. L'abbondanza attuale di deuterio è strettamente connessa alla densità primitiva (*si veda l'illustrazione di questa pagina*). L'abbondanza relativa di deuterio nello spazio interstellare più vicino è stata misurata dal terzo satellite della serie OAO (Osservatorio Astronomico Orbitante), chiamato *Copernicus*. Dopo aver tenuto conto del deuterio prodotto dalle reazioni nucleari nelle stelle, l'abbondanza misurata porta a una densità media attuale di circa 4×10^{-31} grammi per centimetro cubo. Questo tipo di osservazione consente di valutare la densità con buona precisione: se l'universo fosse 10 volte più denso, il big bang avrebbe prodotto meno di un millesimo dell'abbondanza osservata di deuterio. Per questa ragione le incertezze che indubbiamente permangono sulla precisione delle misure non portano a grandi incertezze sulla densità stimata.

Il fatto che la densità determinata dall'abbondanza del deuterio rappresenti un universo aperto o chiuso dipende dal tempo di Hubble. Come abbiamo visto, se il tempo di Hubble è pari a 19 miliardi di anni, la densità critica vale 5×10^{-30} grammi per centimetro cubo, così che Ω, il rapporto tra la densità effettiva e la densità critica, è circa 0,08. Per qualsiasi valore del tempo di Hubble compreso tra 13 e 19 miliardi di anni, il valore di Ω ricavato dall'abbondanza di deuterio è compatibile con quello ottenuto dalla densità delle galassie. Al contrario, per tutti i valori accettabili del tempo di Hubble, un valore di Ω maggiore o circa uguale a 1 non è compatibile con la densità che è necessaria per sintetizzare il deuterio.

L'abbondanza di deuterio sembrerebbe fornire prove particolarmente significative a favore di un universo aperto; purtroppo però gli argomenti che sosten-

L'esame combinato di ulteriori limitazioni indica che l'universo continuerà a espandersi per sempre. L'abbondanza di deuterio pone un limite superiore alla densità di tutta la materia dell'universo, e perciò limita anche il parametro di densità, anche se il valore numerico di tale limite dipende dal tempo di Hubble. Un limite superiore per il tempo di Hubble stesso è fissato dalla stima delle età delle stelle e degli elementi pesanti. Infine, i calcoli sulla massa associata agli ammassi di galassie forniscono un limite inferiore al parametro di densità. Escludendo complicazioni apparentemente poco probabili, i modelli plausibili giacciono entro un piccolo intervallo di valore del parametro di densità e del tempo di Hubble (*superficie colorata*); gli universi corrispondenti sono aperti, infiniti e in espansione perpetua.

gono questa conclusione presentano delle incertezze. Nell'estrapolare dallo stato attuale dell'universo alle condizioni immediatamente successive al big bang si è fatto uso del modello più semplice possibile; altri modelli potrebbero ammettere che le quantità osservate di elio e di deuterio siano state sintetizzate in un universo molto più denso e chiuso. Tali modelli sono più complicati, anche un po' artificiosi, ma non possono essere esclusi. Inoltre, l'importanza dell'abbondanza del deuterio dipende esclusivamente dall'ipotesi che tutto il deuterio presente nell'universo sia stato prodotto subito dopo il big bang. Sono state proposte altre sorgenti, come le supernove, ma finora non è stato trovato nessun meccanismo capace di sintetizzare una quantità significativa di deuterio senza violare altre restrizioni.

Modelli plausibili

Le misure del parametro di decelerazione, dell'età dell'universo, della densità di galassie e dell'abbondanza di deuterio pongono limitazioni indipendenti tra loro sullo stato dell'universo. Se i risultati delle misurazioni non sono contraddittori, deve esistere qualche classe di modelli dell'universo che soddisfi a tutte le condizioni poste. In effetti tale classe esiste ed è anche relativamente piccola, così da permettere interessanti previsioni sul futuro dell'universo (*si veda l'illustrazione di questa pagina*). Se l'universo non è troppo vecchio e se la sua densità è almeno pari a quella osservata nelle galassie, ma contemporaneamente non troppo grande perché sia possibile la produzione del deuterio, il valore di Ω deve essere compreso tra 0,04 e 0,09. È cioè un valore molto minore di quello richiesto per un universo chiuso.

Altre due osservazioni sono compatibili coi valori permessi di Ω e del tempo di Hubble. L'età calcolata per le stelle che formano gli ammassi globulari dipende dall'abbondanza dell'elio, che, come abbiamo visto, dipende a sua volta dalla densità dell'universo. È perciò incoraggiante trovare che l'età dell'universo e l'abbondanza dell'elio che risultano dalle restrizioni combinate sono compatibili con le conoscenze accumulate sulle stelle degli ammassi globulari.

Le limitazioni richiedono anche che il tempo di Hubble stesso sia compreso tra 13 e 20 miliardi di anni. La determinazione diretta del tempo di Hubble è difficile, ma negli ultimi anni Allan R. Sandage e Gustav A. Tammann degli Hale Observatories hanno concentrato i loro sforzi sul problema. Il loro valore più probabile è 18 ± 2 miliardi di anni. Robert P. Kirshner e John Kwan del California Institute of Technology hanno usato una tecnica diversa, che si basa sulle proprietà di stelle che esplodono in galassie lontane, per ottenere un valore indipendente del tempo di Hubble compreso tra 13 e 22 miliardi di anni.

La coincidenza di risultati ottenuti con metodi così diversi è soddisfacente e spinge ad avere fiducia nel modello cosmologico considerato e nel destino dell'universo così previsto. Comunque, a causa delle incertezze nei dati e nella teoria usata per interpretarli, è ancora possibile che questo accordo si riveli fortuito.

Una previsione costante dei modelli qui considerati è che il parametro di decelerazione deve essere uguale alla metà del parametro di densità, e, come abbiamo visto, questa previsione non può ancora essere verificata. Se nel futuro si dovesse trovare che essa è errata, si dovrà fare ricorso a modelli cosmologici più complicati. Per esempio, una classe di modelli fa uso di una modifica della relatività generale suggerita in passato da Einstein, in cui viene introdotto un parametro detto costante cosmologica. In questi modelli è lo spazio stesso che genera una forza gravitazionale attrattiva o repulsiva, e quindi la decelerazione non è più legata in modo semplice alla densità.

Prese una per una tutte le limitazioni discusse permettono interpretazioni alternative. In particolare, alcuni nostri colleghi non concorderebbero con la piccola densità derivata dalla stima della massa associata alle galassie, né con l'inserimento di una limitazione della densità basata sulla produzione di deuterio. I nostri argomenti e le nostre conclusioni poggiano però la loro credibilità sul fatto che è possibile costruire un modello cosmologico interpretando ogni singola indicazione nel modo più semplice. È significativo che fattori così diversi come l'età delle stelle, la massa delle galassie, l'abbondanza degli elementi chimici e la velocità di espansione dell'universo osservata ricevano tutti un'interpretazione naturale all'interno di uno dei più semplici modelli cosmologici. Questo modello descrive un universo di estensione infinita che continuerà a espandersi per sempre. L'argomentazione a favore di un universo aperto non è certo inoppugnabile, ma è fortemente sostenuta dal peso dell'evidenza sperimentale.

Le risorse della percezione binoculare

Studi con stereogrammi che cambiano a caso rivelano che il sistema percettivo estrae dai dati visivi le informazioni sulla profondità e il movimento ancora prima che noi siamo consci di ciò che vediamo

di John Ross

Quando guardiamo con entrambi gli occhi i vari oggetti dello spazio tridimensionale, il nostro sistema visivo non si limita a fondere semplicemente le due immagini retiniche, ma fa molto di più. È noto che la percezione binoculare della profondità dipende dalle differenze tra queste due immagini. Essa, oltre a dimostrarsi molto sensibile alle differenze relative di posizione, esibisce una notevole attitudine a prendere decisioni. Il sistema visivo, in effetti, costruisce le scene tridimensionali partendo dalle immagini bidimensionali formatesi sulle retine, adattando queste informazioni visive in uno schema concettuale. Gli interrogativi che possiamo proporci al riguardo sono numerosi; tra i più importanti ne ricordiamo due: quali capacità di elaborare i messaggi si trovano nella percezione binoculare? Quali sono le risorse a disposizione di queste capacità di elaborare l'informazione?

Con l'invenzione dello stereoscopio, fatta da Charles Wheatstone nel 1938, divenne possibile avere risultati tridimensionali con due fotografie di una scena scattata da punti lievemente diversi. Lo stereoscopio presenta una fotografia all'occhio destro, e l'altra all'occhio sinistro. Quando le due immagini sono guardate contemporaneamente, esse si combinano e formano una scena tridimensionale. Dato che ciascun occhio vede un'immagine completa, i primi studiosi della percezione binoculare della profondità credettero che questa risultasse dal combinarsi delle due immagini monoculari (o, almeno, che dipendesse dal riconoscimento delle caratteristiche percepibili anche con un solo occhio).

Altri studiosi, invece, non accettarono l'ipotesi della combinazione delle singole immagini. Nel 1887 A.L. Austin di Invercargill, in Nuova Zelanda, comunicò a Charles Darwin una curiosa scoperta che aveva fatto: «Per quanto lei non mi conosca, e per quanto viva sull'altra faccia della Terra, mi prendo la libertà di scriverle di una mia piccola scoperta riguardo la visione binoculare nello stereoscopio. Se prendiamo due comuni fotografie formato tessera con due volti diversi (i ritratti siano circa delle stesse dimensioni) e le poniamo in uno stereoscopio, si ha la fusione del tutto straordinaria in un volto unico; con i tratti di alcune signore, in ogni caso, si produce un *deciso miglioramento* della bellezza».

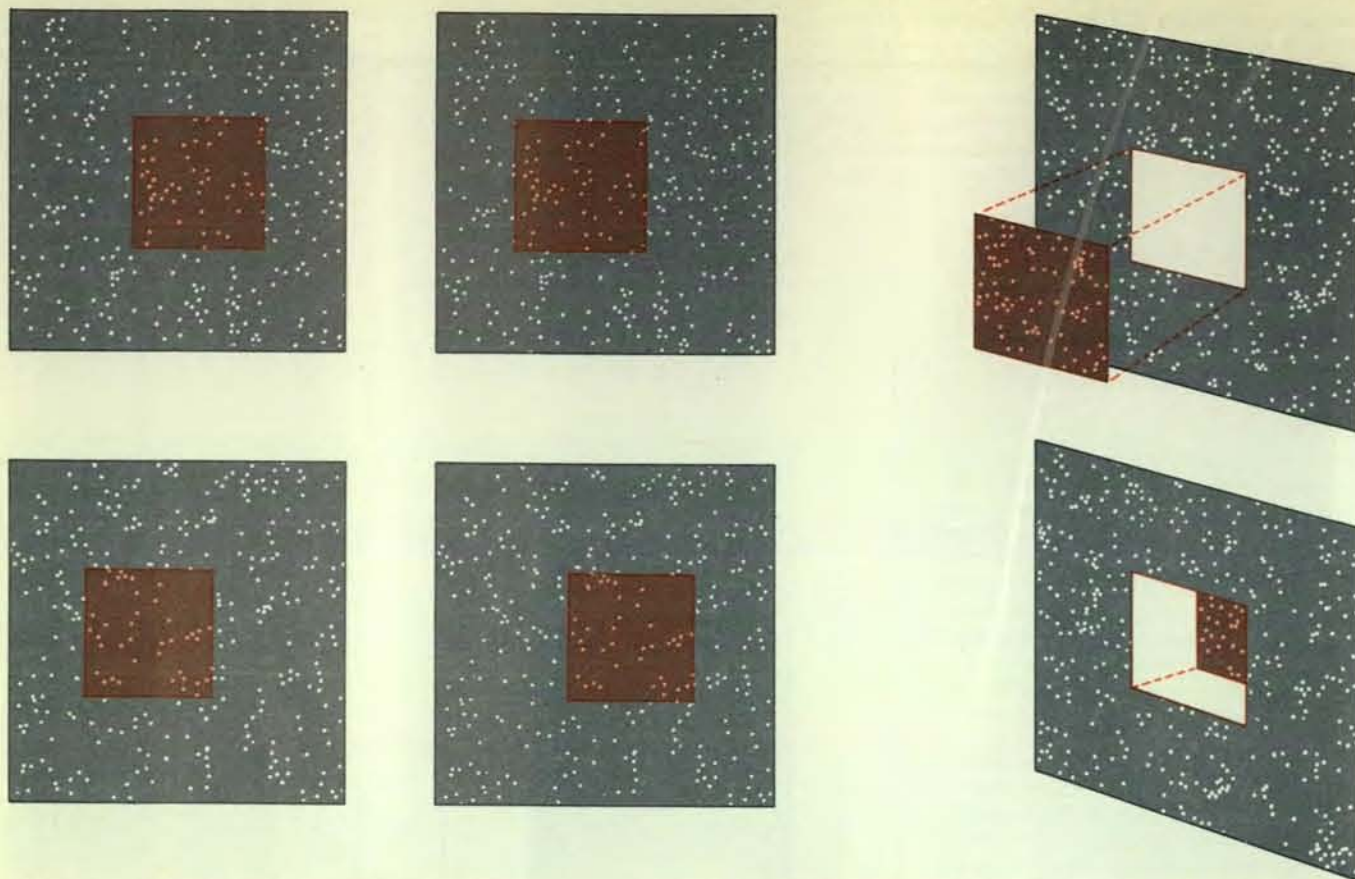
Darwin passò la lettera a Francis Galton, il quale confermò le osservazioni di Austin. Galton era anche a conoscenza di un matematico piuttosto eccentrico che aveva combinato, in uno stereoscopio, due sue fotografie: «una di esse - riferì Galton - lo riproduceva con un'aria austera, l'altra con un sorriso, e questa combinazione portava a una curiosa e forte fusione delle due fotografie». Galton non era tuttavia convinto che questo fenomeno potesse essere spiegato con la combinazione delle singole immagini; egli pensava infatti che la percezione binoculare di due volti differenti dovesse essere diversa da una vera combinazione ottica dei volti.

Possiamo anche avere un effetto analogo quando due volti reali sono guardati con un sistema di specchi che dirige l'immagine di un volto all'occhio sinistro, e quella dell'altro all'occhio destro (*si veda l'illustrazione nella pagina a fronte*). Quanto vorrei fare notare a questo punto è che la combinazione binoculare rivela qualche facoltà critica nel sistema visivo che, apparentemente su basi estetiche, è capace di prendere decisioni e di respingere informazioni.

I vantaggi della percezione binoculare sono stati dimostrati nella loro interezza da Bela Julesz, dei Bell Laboratories, con i suoi stereogrammi con punti disposti a caso. Per mezzo di un calcolatore egli ha creato degli insiemi casuali di punti che, guardati con uno stereoscopio, si combinano per formare delle scene tridimensionali. Quando i medesimi stereogrammi sono invece guardati con un occhio, si vede una tessitura distribuita completamente a caso, senza che si riesca a cogliere delle forme particolari. Julesz ha dimostrato in modo decisivo che è possibile avere percezione binoculare della profondità anche senza indizi monoculari riconoscibili.

Nel nostro laboratorio alla University of Western Australia, John H. Hogben e io, in stretta collaborazione con Monte

Combinando binocularmente due volti diversi si ottiene una curiosa fusione, spesso con un chiaro miglioramento delle loro caratteristiche. Gli specchi sono disposti in modo che l'occhio sinistro vede il volto di sinistra e l'occhio destro quello di destra. Quando un volto è quello di un uomo e l'altro è quello di una donna si ottengono effetti fuori dal comune. Questa fusione selettiva delle caratteristiche, propria della combinazione binoculare, indica che il sistema visivo ha evidentemente la capacità di accettare o di rifiutare informazioni su basi estetiche.

In questa illustrazione è mostrato, schematicamente, un sistema di generazione di stereogrammi con punti disposti completamente a caso mediante calcolatore. Un generatore di numeri casuali emette le coordinate per i punti luminosi da presentare, e un calcolatore aggiunge delle informazioni di profondità al flusso continuo di punti. I punti compaiono brevemente in coppie sugli oscilloscopi. Per ciascun punto sullo schermo di sinistra c'è un punto corrispondente sullo schermo di destra. L'osservatore vede migliaia di punti luminosi che compaiono e scompaiono a caso. Nonostante il fatto che la scena cambi di continuo, il sistema visivo è capace di mettere assieme le coppie di punti dei due schermi. L'osservatore percepisce il tutto come una singola scena avente una chiara profondità, con oggetti che si ergono liberi nello spazio.

Le informazioni spaziali sono aggiunte per mezzo di uno spostamento orizzontale dei punti appartenenti alle regioni che debbono essere viste in profondità. Per creare un quadrato centrale che si solleva, ad esempio, i punti sono spostati a destra nell'immagine di sinistra e a sinistra nell'immagine di destra (*in alto*). Dove si ha una sovrapposizione di punti, i punti dello sfondo sono eliminati dal calcolatore. Nelle aree vuote che si hanno come conseguenza dello spostamento, invece, sono aggiunti punti disposti a caso. Per creare l'effetto di una superficie più lontana, i punti sono spostati a sinistra nell'immagine di sinistra e a destra nell'immagine di destra (*in basso*).

Sala, un ingegnere elettronico, abbiamo elaborato un sistema basato su calcolatori aventi maggiori possibilità di quello di Julesz, ma sempre con stereogrammi di punti disposti a caso. Il terminale del nostro sistema consiste di un paio di oscilloscopi in cui possono venire presentati dei punti e che sono otticamente separati: uno è quindi visto dall'occhio destro e l'altro dall'occhio sinistro. I punti luminosi sono esposti per brevi tempi e in coppia, uno sull'oscilloscopio di destra e uno su quello di sinistra. Ciascun occhio dell'osservatore vede una rapida successione di punti luminosi, ognuno dei quali è indipendente dai punti precedenti e da quelli successivi. Quando l'osservatore chiude un occhio, tutto quello che vede è uno sciame di punti luminosi che sembrano muoversi sull'intero schermo quadrato, come la «neve» che appare sullo schermo di un comune televisore.

La sequenza dei punti luminosi non è preparata in anticipo, né viene ripetuta. Un generatore di numeri casuali progettato a questo scopo produce coppie di coordinate per i punti che vengono disposti sulla griglia 256 × 256 degli schermi degli oscilloscopi. Ciascun punto luminoso è così generato a caso.

A questo flusso continuo di punti vengono aggiunte delle informazioni di profondità per mezzo di un calcolatore che controlla quanto viene emesso. Il calcolatore ha un programma della scena da rappresentare. Appena viene intercettato un punto, il calcolatore determina l'eventuale spostamento nella sua posizione. I principi alla base di questo spostamento sono gli stessi della stereoscopia classica. I punti che appartengono a una superficie vicina sono spostati a destra nel campo dell'occhio sinistro. Per rappresentare i punti di una superficie lontana si fa il contrario. Quanto è maggiore lo spostamento dei punti, tanto è maggiore la profondità che viene percepita. Per assicurare una distribuzione uniforme di punti nelle regioni in cui lo spostamento crea una sovrapposizione, il calcolatore elimina i punti dello sfondo. Nelle regioni in cui, di contro, lo spostamento crea un vuoto, il calcolatore aggiunge dei punti.

Consideriamo ciò che deve fare il sistema visivo per giungere, in queste circostanze, a un percetto binoculare. Migliaia di punti compaiono e scompaiono a caso. Per ciascun punto visto da un occhio, l'altro occhio deve scegliere il punto con il quale è accoppiato. Inoltre, due punti che abbiano subito uno spostamento nelle posizioni reciproche debbono essere interpretati come un singolo punto, non come due punti differenti visti da occhi differenti. Infine, se una superficie deve essere vista in profondità come una forma con contorni ben definiti, il sistema visivo deve essere in grado di riconoscere tutti i punti appartenenti a quella forma come aventi una disparità comune.

Un osservatore con visione stereoscopica normale, dopo un breve periodo di adattamento, vede quanto avviene sugli oscilloscopi come una scena in profondità. Con la pratica, l'intervallo di tempo richiesto per arrivare a un percetto tridimensionale diminuisce fino ad annullarsi. Quando si vede per la prima volta una scena prodotta da stereogrammi con punti disposti a caso si prova una sensazione strana. Gli oggetti si ergono liberi nello spazio, ben chiari e delimitati. Vediamo un semplice caso. Una regione quadrata al centro del video è spostata dal calcolatore in modo tale che nell'oscilloscopio di destra i punti del quadrato centrale sono spostati a sinistra, e nell'oscilloscopio di sinistra sono spostati a destra. Quando l'osservatore guarda con un occhio uno degli insiemi stimolanti, vede solo una massa di punti luminosi che si muovono su uno sfondo scuro. Quando guarda entrambi gli insiemi con i due occhi, invece, la scena cambia in modo spettacolare. Al centro dell'immagine si vede un quadrato che ondeggia davanti allo sfondo. Il quadrato è come un pezzo di plastica scura sul quale vanno e vengono i punti luminosi.

Per quanto l'insieme dei punti della superficie muti di continuo, il quadrato dà l'impressione di essere solido e immutevole. Il quadrato solido, in realtà, non esiste, ma è stato costruito dalla percezione binoculare per dare una spiegazione alle informazioni che sta ricevendo dai punti disposti a caso degli stereogrammi.

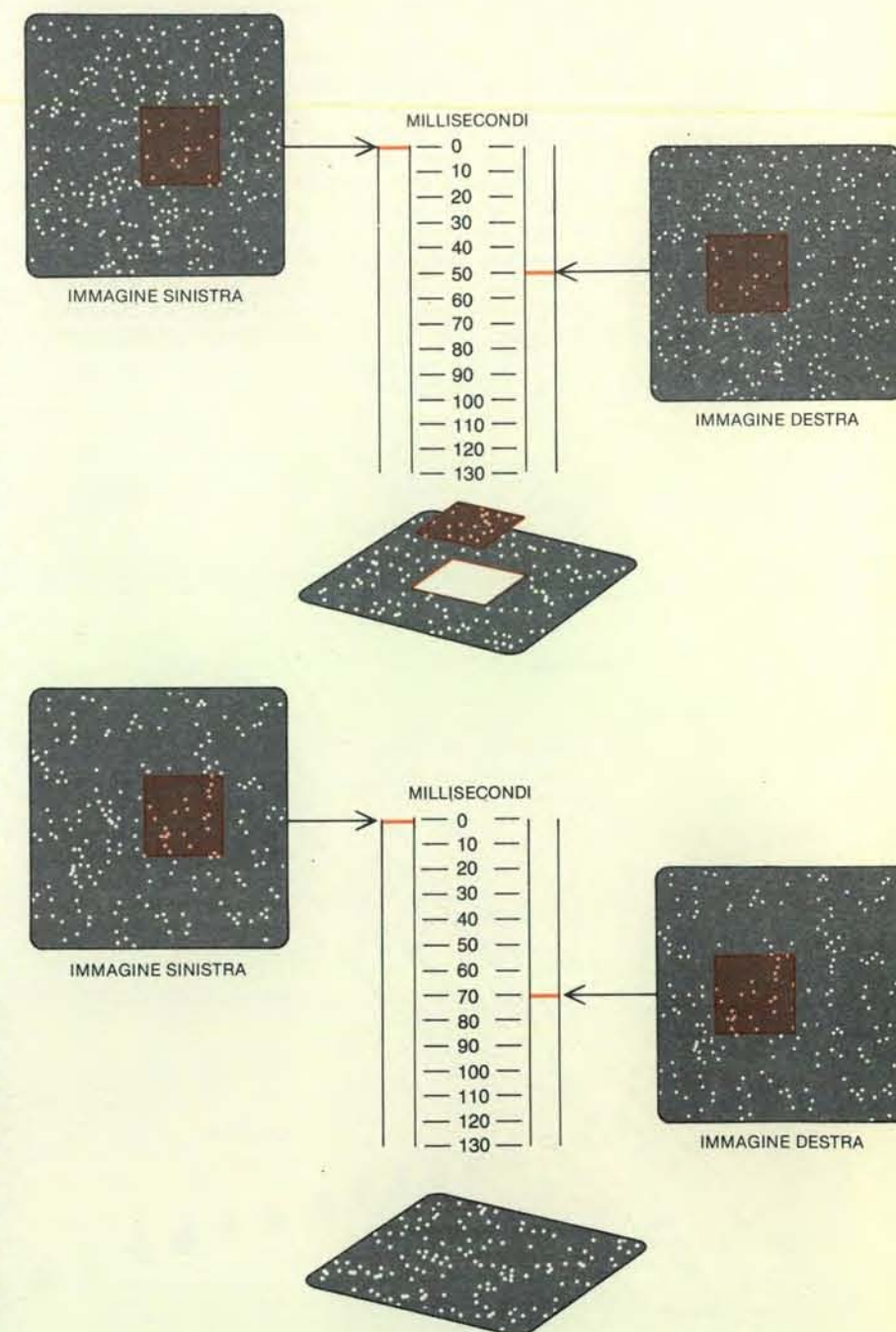
C'è qualcos'altro, nella scena binoculare, che colpisce l'osservatore dopo un momento di riflessione: i punti luminosi entro al piccolo quadrato sono come bloccati sulla sua superficie. Va ricordato che quando si guarda il tutto con un occhio, i punti si spostano sullo schermo in modo completamente uniforme. Ora, con la visione binoculare, alcuni punti sono sollevati rispetto allo sfondo e non oltrepassano mai il confine del loro nuovo territorio; sembra invece che, arrivati agli orli del quadrato, rimbalzino. Ciò significa che il sistema visivo attribuisce un significato funzionale agli orli costruiti dal processo binoculare, e che è impedito lo spostamento attraverso questi orli: non si può cioè «saltare» da un livello di profondità a un altro. Inoltre, nel caso in cui la percezione monoculare entri in conflitto con i costrutti binoculari, i percetti monoculari vengono soppressi.

Nei nostri esperimenti non si è trovato un limite superiore alla velocità con cui la percezione binoculare riesce ad affrontare l'entrata delle coppie di punti. Il nostro calcolatore può produrre fino a 30 000 coppie di punti al secondo, velocità che è affrontata facilmente dal sistema visivo. Noi abbiamo un sistema ottico che arriva a emettere 250 000 coppie di punti al secondo, e anche con questi valori la percezione binoculare non dà segno di essere caricata da un numero di informazioni eccessivo per le sue possibilità. È ovvio che un limite superiore deve esserci, perché quando la velocità di emissione è sufficientemente elevata lo schermo viene inondato di luce. Anche in questo caso, la limitazione può trovarsi a livello dei recettori retinici, più che a livello delle capacità di elaborare i messaggi da parte dei meccanismi che servono a confrontare l'entrata dei due occhi.

C'è, invece, un limite inferiore al di sotto del quale l'informazione giunge a una velocità troppo bassa per mantenere la percezione di forme in profondità. Tale limite varia a seconda della complessità della scena presentata e della grandezza dei dettagli della scena. Per singole forme di grandezza ragionevole il limite è di circa 2000 coppie di punti al secondo. Al di sotto di questo valore i punti possono ancora sembrare a distanze diverse, ma non è più visibile la forma e non ci sono contorni ben definiti.

La percezione binoculare, degli stereogrammi con punti distribuiti a caso, è capace di costruire scene molto complesse: è solo necessario che la velocità di emissione dei dati sia abbastanza elevata per mantenere la scena. Si può allora facilmente percepire fino a una stratificazione di 10 piani posti a distanze diverse. Il calcolatore ha anche la possibilità di presentare stimoli che sono visti come superfici inclinate.

Nel caso in cui siano programmate due superfici in una medesima zona, ci si potrebbe aspettare di avere il caos, ma non è così. Metà dei punti luminosi sono visti su una superficie anteriore e metà su una superficie che sta dietro. Ecco quindi risolto, e in modo elegante, il problema di rendere visibili entrambe le superfici: la superficie anteriore sembra trasparente e l'altra sembra opaca. Ciò dimostra che la percezione binoculare riesce a cogliere i rapporti spaziali in modo intuitivo e tale da trascendere la pura e semplice geometria.

Come si è detto, i costrutti della percezione binoculare controllano il movimento apparente impedendo di oltrepassare i contorni e di saltare in regioni con profondità diverse. Ci siamo allora chiesti: è possibile far sì che una figura sembri avvicinarsi o allontanarsi rispetto all'osservatore? Abbiamo modificato il programma del calcolatore in modo tale che la disparità dei punti sulla superficie della figura cambiasse continuamente. Ammesso che il cambiamento non sia troppo rapido o troppo lento, l'osservatore percepisce effettivamente una figura che si

Le limitazioni temporali della percezione binoculare si sono studiate anche presentando il flusso di punti all'occhio sinistro in un dato momento, e all'occhio destro con un certo ritardo. Quando il ritardo tra le due immagini è inferiore ai 50 millisecondi (*in alto*), le figure sono viste in profondità. Quando il ritardo supera i 50 millisecondi, l'impressione di profondità viene meno (*in basso*).

muove senza alterare la sua struttura. La figura cioè mantiene la forma e i suoi contorni rimangono solidi, continuando a controllare e a bloccare il movimento apparente dei punti contenuti nella figura stessa. Con un movimento in avanti e in dietro della figura non si perde quindi la percezione della profondità.

C'è un limite alla disparità spaziale che può essere introdotta negli stereogrammi. Una disparità troppo grande dà luogo a immagini doppie, per quanto non necessariamente a una caduta completa del senso di profondità. Hogben e io abbiamo voluto determinare se ci fosse stato un limite temporale, oltre a quello spaziale. In altre parole, potevamo presentare una serie di punti a un occhio e, un poco più tardi, la serie di punti corrispondenti all'altro occhio continuando ad avere una fusione binoculare?

A tale fine abbiamo introdotto un ritardo alla serie di punti, sempre disposti a caso, destinati a uno degli occhi. I punti per l'occhio sinistro venivano quindi presentati in un certo momento, mentre quelli per l'occhio destro venivano presentati successivamente. Entrambi gli oscilloscopi ricevevano punti con la medesima velocità ed entrambi gli schermi, guardati monocularmente, sembravano identici; il flusso di punti su uno schermo veniva però dopo, come se fosse partito più tardi.

Consideriamo il problema che la percezione binoculare, in queste circostanze, si trova di fronte. Due punti giungono nello stesso momento, uno per ciascun occhio, ma sono completamente indipendenti l'uno dall'altro, e non portano quindi informazioni sulla disparità. Altri punti arrivano alla velocità di 10 al millisecondo. Sarebbe impossibile mettere assieme le coppie di punti, a meno che non si abbia un qualche tipo di registrazione delle migliaia di punti visti. Per di più, questa registrazione deve essere estremamente fine se vogliamo utilizzarla per trovare l'esatta disparità tra coppie di punti.

Secondo i nostri risultati, la percezione binoculare può tollerare un ritardo di circa 50 millisecondi (un ventesimo di secondo), ma non di più. Con un poco di pratica gli osservatori riescono a vedere delle forme in profondità e a identificarle fino a quando il ritardo tra i punti è al di sotto dei 50 millisecondi. Se è più lungo anche di pochi millisecondi, l'impressione di profondità viene meno e non si vede più alcuna forma. Il ritardo tollerabile varia un poco, da soggetto a soggetto, intorno al valore dei 50 millisecondi, ma è costante per ogni soggetto.

Dai risultati di questo esperimento possiamo trarre due importanti conclusioni. La prima è che la percezione binoculare è soggetta a limitazioni sia temporali sia spaziali. La disparità temporale deve essere contenuta entro un intervallo di 50 millisecondi, così come la disparità spaziale non deve oltrepassare una certa distanza. La seconda conclusione, che si può trarre dall'esperimento, è che la percezione binoculare deve incorporare un qualche genere di memoria visiva capace di mantenere, per almeno 50 millisecondi, una fine registrazione delle posizioni di migliaia di punti.

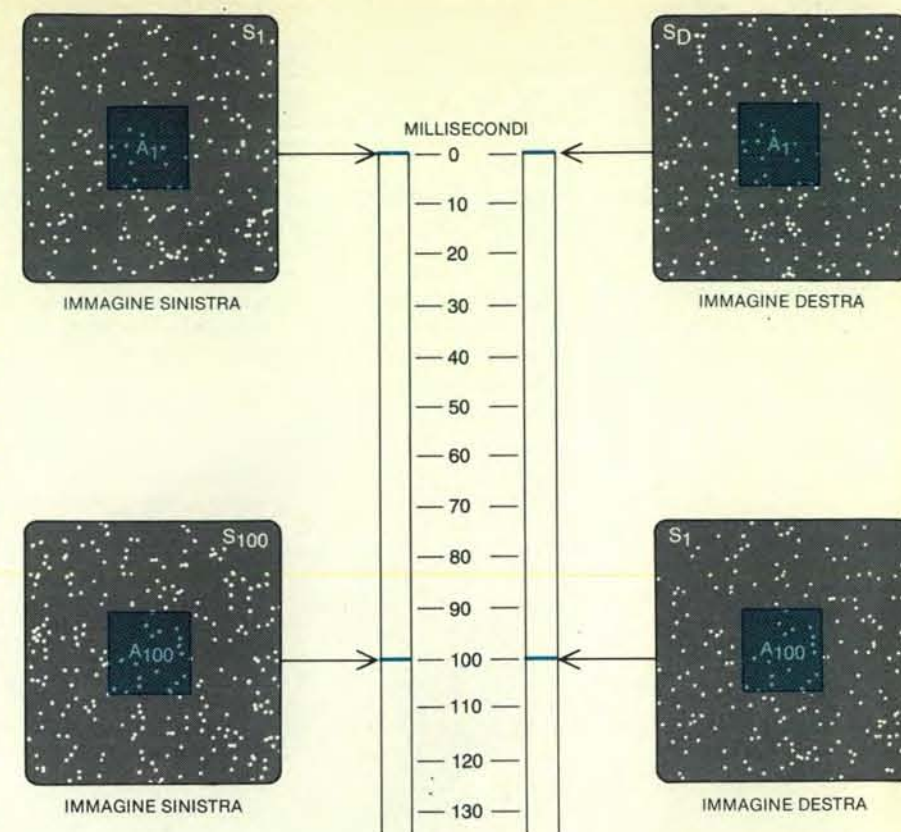
Nelle condizioni del nostro esperimento, tuttavia, per poco che vediamo un punto luminoso, esso rimane visibile per circa 130 millisecondi. Questo valore può essere determinato in vari modi: per esempio, contando semplicemente tutti i punti visibili in ogni istante su una piccola porzione dello schermo. Il problema sta quindi nello spiegare perché il limite di tempo per il confronto binoculare è più breve del tempo in cui il punto rimane visibile. Se un punto è ancora visto da un occhio, perché non riusciamo ad accoppiarlo a un secondo punto che compare all'altro occhio? Prima di passare a considerare questo paradosso, vediamo qual è il ruolo della percezione binoculare nel seguire oggetti che si spostano lateralmente.

Quando si segue con lo sguardo un oggetto in movimento, si creano differenze temporali nella entrata di informazioni visive ai due occhi. Se gli occhi convergono su un oggetto, per esempio un uccello che si sposta verso destra, l'occhio sinistro vede gli elementi dello sfondo prima dell'occhio destro. Per esempio, nel momento t_1 l'occhio sinistro vedrà l'albero che si trova nello sfondo, ma l'occhio destro non lo vedrà che nel momento t_2. L'occhio sinistro, nel frattempo, si è spostato in avanti e ora è puntato sul palo del telefono. Nel momento t_3 l'occhio destro incontra il palo, mentre l'occhio sinistro vede la casa. La situazione risulta invertita per le cose che si trovano in primo piano.

Quando fissiamo gli occhi su un oggetto, questo forma immagini in posizioni corrispondenti delle due retine. Gli oggetti più lontani o più vicini di quello fissato formano immagini in posizioni differenti per ciascuna retina, dando così luogo a quella disparità che sta alla base della percezione stereoscopica della profondità. Quindi, in qualunque momento gli occhi hanno un'immagine differente sia dello sfondo sia del primo piano, ma un'immagine identica dell'oggetto fissato. Se l'oggetto si sposta e lo seguiamo con lo sguardo, ciascun occhio mantiene la medesima immagine dell'oggetto, ma ora incontra anche gli stessi elementi dello sfondo e del primo piano in momenti diversi (*si veda l'illustrazione nella pagina a fronte*). Se lo sguardo segue un oggetto che si sposta verso destra, l'occhio sinistro incontrerà ciascuna parte dello sfondo prima dell'occhio destro. La situazione si inverte per quanto riguarda il primo piano: l'occhio destro, cioè, incontrerà ciascun particolare del primo piano prima dell'occhio sinistro.

Sono riuscito a riprodurre, con gli stereogrammi di punti disposti a caso, l'informazione visiva che si ha quando seguiamo con lo sguardo un oggetto mobile. Oltre a tutto, questi stereogrammi eliminano qualsiasi indizio monoculare riguardante sia l'oggetto sia lo sfondo o il primo piano. Al fine di determinare se le differenze temporali che si verificano nel seguire un oggetto in movimento hanno un ruolo nella percezione binoculare, il calcolatore era programmato per presentare la figura esattamente nello stesso momento su entrambi gli schermi. I punti al di fuori di questa figura erano presentati a un occhio immediatamente, e all'altro con un piccolo ritardo (*si veda l'illustrazione in questa pagina*). Non vi è disparità spaziale negli stereogrammi. Tutti i punti sono ordinati esattamente nella stessa posizione su entrambi gli schermi.

L'osservatore vede la figura in profondità, cosa che non ci sorprende dato che l'area della figura è presentata contemporaneamente a entrambi gli occhi, mentre lo sfondo non lo è. Quanto ci sorprende è che l'area attorno alla figura è percepita come sfondo o come primo piano, e che nel primo caso si sposta in una direzione mentre nel secondo caso si sposta nella direzione opposta.

Se i punti luminosi della zona esterna alla figura raggiungono prima l'occhio sinistro, lo sfondo si sposta da destra verso sinistra, come se la figura scivolasse sullo sfondo andando da sinistra verso destra. Se i punti raggiungono prima l'occhio destro, lo sfondo si sposta da sinistra verso destra. In entrambi i casi la parte che sembra più vicina a noi, cioè il primo piano, si muove in direzione opposta a quella in cui si muove lo sfondo. Effettivamente, sfondo e primo piano possono combinarsi nel dare la forte impressione di un cilindro in piedi che ruota attorno al suo asse verticale, mentre la nostra figura si trova nel mezzo del cilindro.

Le differenze temporali nell'entrata d'informazioni che si hanno quando gli occhi seguono un oggetto mobile possono essere riprodotte per mezzo di stereogrammi dinamici con punti disposti a caso. I punti luminosi dell'area centrale sono presentati contemporaneamente a entrambi gli occhi, ma l'occhio destro vede i punti dell'area circostante dopo un piccolo ritardo. Per esempio, l'occhio sinistro vede A_1 ed S_1 mentre quello destro vede A_1 e una tessitura circostante differente (S_D). Cento millisecondi più tardi l'occhio sinistro vede A_{100} ed S_{100} mentre quello destro vede A_{100} ed S_1. Non c'è disparità spaziale: entrambi gli occhi vedono tutti i punti esattamente nella stessa posizione. Il ritardo temporale tra i due occhi, tuttavia, crea un effetto di profondità e di movimento analogo a quello illustrato nella figura della pagina successiva.

Un effetto analogo possiamo ottenerlo anche, molto alla buona, sintonizzando il televisore su un canale privo di programmi e adattando il contrasto in modo da avere un buon «effetto neve». Mettiamo poi davanti a un occhio un filtro scuro o una lente per occhiali da sole, e guardiamo lo schermo con entrambi gli occhi. Il filtro o la lente degli occhiali ritarda l'informazione ricevuta da quell'occhio, e riusciremo così ad avere l'effetto del cilindro che ruota. La neve andrà in una direzione di fronte allo schermo e, dietro di questo, nella direzione contraria. Quando mettiamo la lente scura davanti all'altro occhio, la direzione del movimento si inverte.

Nel nostro esperimento c'erano due indicazioni di profondità in conflitto tra di loro: la disparità che deponeva per una scena statica, e il ritardo che deponeva per un movimento continuo. Con ritardi al di sotto dei 50 millisecondi si vede la scena statica, ma con ritardi al di sopra di 70 millisecondi si aveva un risultato di movimento. Ritardi tra i 50 e i 70 millisecondi portavano a continue oscillazioni tra un effetto di movimento e un'impressione di staticità.

I risultati dimostrano che la percezione binoculare può sintonizzarsi con diverse fonti di informazione presenti in una data situazione. Per ritardi al di sotto dei 50 millisecondi si impone una disparità spaziale. Per ritardi di 70 millisecondi e, riteniamo, fino a 2 secondi, le differenze temporali danno luogo alla percezione di un movimento continuo e in profondità.

In altre circostanze, ritardi inferiori al millisecondo, dell'ordine dei 160 microsecondi, sono sufficienti per dare una chiara indicazione di profondità. David Burr e io lo abbiamo dimostrato con delle sequenze stroboscopiche che danno l'illusione di un movimento omogeneo. Ciascun occhio vede la medesima sequenza, ma c'è una differenza di fase tra le sequenze osservate dai due occhi. Questa differenza di fase è colta dall'osservatore e interpretata come una indicazione di profondità.

Ora riteniamo che la visione abbia due modi ugualmente efficaci di trattare gli oggetti in movimento. Possiamo cogliere la disparità spaziale istantanea per determinare la profondità, o possiamo cogliere la differenza di fase con cui gli oggetti, nei due occhi, si muovono nei confronti di punti di riferimento comuni. Vale a dire che la visione, nei confronti delle informazioni che riceve, può adot-

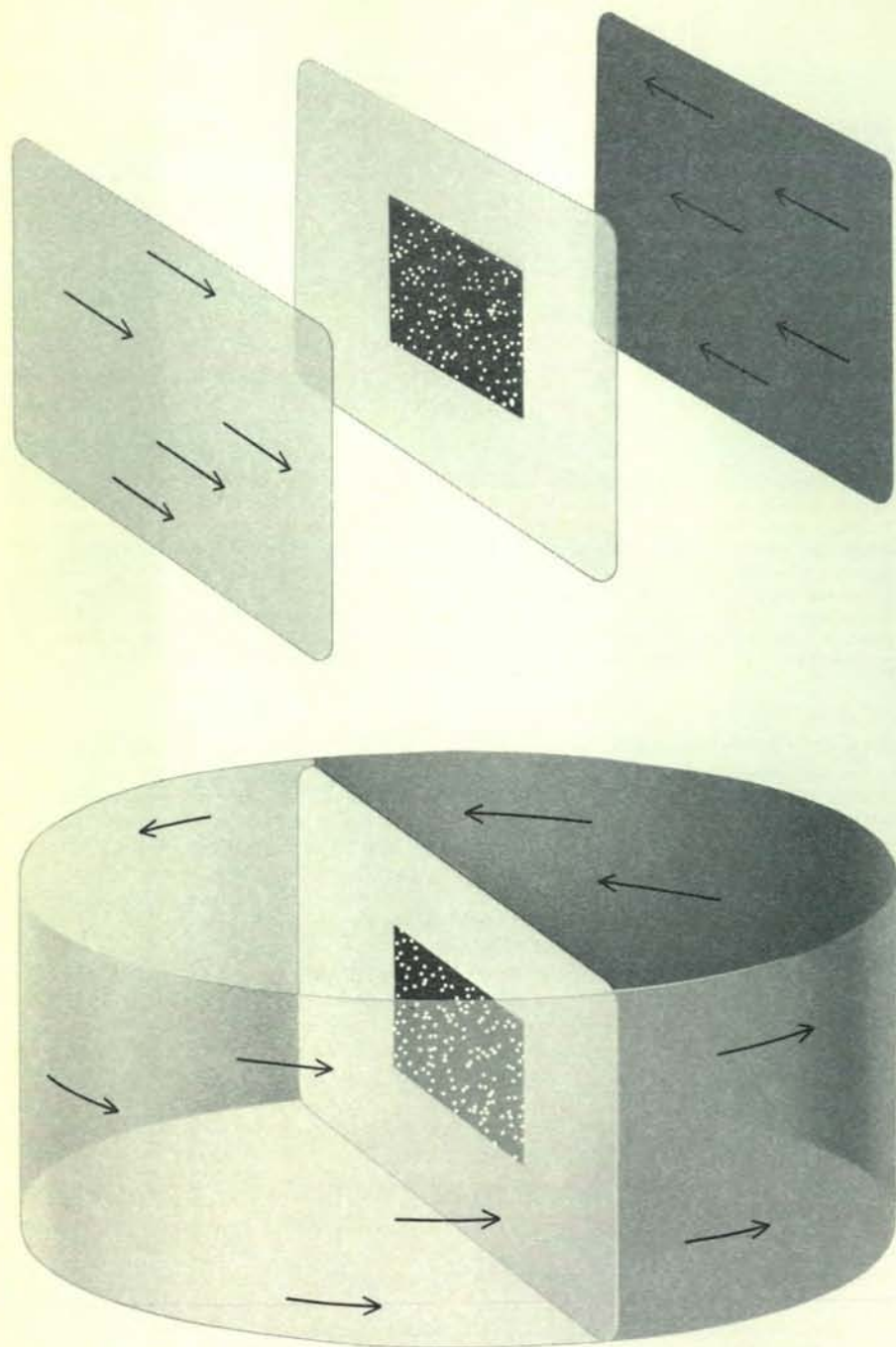
tare uno di due atteggiamenti percettivi. Nel mondo reale i percetti che risultano da questi diversi atteggiamenti sono generalmente coerenti tra di loro, ma in condizioni sperimentali, come la nostra, la scena può variare a seconda dell'atteggiamento.

Torniamo ora al paradosso per cui un punto luminoso che rimane visibile per 130 millisecondi non può essere accoppiato con un secondo punto, se questo compare con un ritardo di più di 50 millisecondi. Evidentemente, in questo caso possiamo vedere qualcosa, ma non riusciamo a utilizzare le informazioni. D'altro lato, quando dei ritardi temporali più lunghi portano a vedere un movimento, stiamo ovviamente utilizzando delle informazioni su una massa di punti che sono già scomparsi. Ciò significa che deve essere mantenuto un qualche genere di registrazione visiva per un tempo maggiore di 130 millisecondi. In parole diverse, quando si segue con lo sguardo un movimento laterale, le registrazioni delle informazioni in entrata sono mantenute oltre i limiti di tempo della visibilità.

Ritengo si debba concludere che la percezione binoculare ha accesso a registrazioni di dati visivi che sono indipendenti da ciò che vediamo. Questo rappresenta una rottura radicale con quanto si ritiene comunemente: e cioè che quello che vediamo sia alla base di tutti i dati sensoriali sui quali i processi percettivi superiori elaboreranno poi una idea della scena che abbiamo davanti agli occhi. Sembra che vi siano delle registrazioni di dati visivi che si possono consultare prima che riusciamo a vedere qualsiasi cosa, e ciò al fine di stabilire uno schema appropriato entro cui fare entrare il risultato percettivo.

Quando abbiamo di fronte stereogrammi con punti disposti a caso, ciascun occhio guarda uno stereogramma distinto ma non «vediamo» due stereogrammi. Il nostro percetto binoculare è il risultato delle differenze spaziali e temporali presenti negli stereogrammi. Le fini discrepanze tra le due vedute debbono essere registrate separatamente e molto accuratamente. Non sappiamo né cosa siano queste registrazioni, né quanti memorizzatori separati vi siano, né quale sia la loro base fisica. Alcuni fenomeni visivi particolarmente problematici (come il movimento apparente, il mascheramento visivo e la percezione della simmetria) possono essere considerati assieme ammettendo che la percezione, prima che noi siamo in grado di percepire, deve servirsi dell'analisi di registrazioni visive indipendenti.

Alcune capacità della percezione binoculare (per esempio, allineare due immagini di grandezza diversa o riconoscere che un'intera regione ha un valore comune di disparità) sono spiegate assai elegantemente dal modello della percezione stereoscopica proposta da Julesz. Ma altre capacità di selezione, di analisi e di sintesi (come la percezione di un volto idealizzato), e quelle che ci fanno cogliere le differenze temporali debbono ancora essere spiegate.

Un indirizzo lo possiamo avere nel fatto che le forme rese visibili dalla combinazione binoculare di punti ordinati a caso sono maggiormente idealizzate delle forme reali. Le configurazioni quadrate sono quadrati più perfetti, e con contorni più perfetti di qualsiasi quadrato reale. Sono come la forma platonica di un quadrato. Ciò che osserviamo in questi stereogrammi con punti distribuiti a caso possono bene essere delle concezioni idealizzate, imposte al flusso esterno di informazioni da un qualcosa che è nel nostro sistema visivo. Ciò che osserviamo può essere una struttura imposta dalla visione nel momento in cui la si sintonizza con le fonti di informazione che riesce a cogliere. Così come un calcolatore ha il suo programma, il sistema visivo può avere un programma di adattamenti per le forme nello spazio e nel tempo. Quello che vediamo è un'interpretazione del mondo esterno, ordinata entro uno schema di riferimento che il sistema visivo impone a causa dell'atteggiamento che adotta. In altri termini, per comprendere il mondo noi adottiamo un atteggiamento percettivo.

Quando negli stereogrammi con punti disposti a caso sono introdotti dei ritardi temporali, si ottiene un effetto di profondità e di movimento (come si è già accennato nella illustrazione della pagina precedente). La figura è vista come un oggetto solido. L'area circostante è tuttavia percepita sia come sfondo sia come primo piano (*in alto*). I punti luminosi del primo piano si spostano in una direzione e quelli dello sfondo si spostano nella direzione opposta, facendo sembrare che la figura si sposti verso destra. Il primo piano e lo sfondo possono combinarsi per dare l'impressione di un cilindro in posizione verticale che ruota attorno alla figura (*in basso*).

I piccoli calcolatori elettronici

L'elemento di base dei calcolatori tascabili è un chip microelettronico. I circuiti realizzati sul chip e i componenti a essi associati formano un sistema di elaborazione dell'informazione notevolmente sofisticato

di Eugene W. McWhorter

Nei cinque anni trascorsi dall'invasione del mercato da parte dei calcolatori elettronici tascabili milioni di persone hanno avuto la gradevole esperienza di vedere, mostrato dall'indicatore, il risultato di un'operazione come 953,22 per 14,331 nello stesso istante in cui era stato premuto il tasto «uguale». Sembra però quanto meno prudente affermare che solo pochi di coloro che hanno apprezzato una tale abbreviazione dei tempi ed espansione delle capacità di calcolo possiedono più di una vaga conoscenza di ciò che accade all'interno del calcolatore. È risaputo che i processi che si svolgono in un calcolatore elettronico tascabile si compiono in un piccolo *chip* microelettronico (cioè in una piastrina che per la sua piccolezza viene chiamata appunto «chip» ovvero «scheggia»), ma non è altrettanto noto come l'organizzazione logica e le procedure numeriche siano collegate con questi processi.

Sia per quanto riguarda le origini, sia per i principi operativi il piccolo calcolatore elettronico è uno sviluppo diretto dei calcolatori elettronici «da ufficio» e «professionali da tavolo». Circa dieci anni fa un tipico calcolatore a quattro funzioni (addizione, sottrazione, moltiplicazione e divisione) di quel genere incorporava centinaia di circuiti integrati microelettronici separati e costava diverse centinaia di dollari. Alcuni lungimiranti tecnici dell'industria elettronica prevedevano tuttavia che un giorno gli stessi concetti sarebbero stati incorporati in macchine tanto piccole, semplici ed economiche da riuscire a conquistare un mercato di massa. Già nel 1965 presso la Texas Instruments Incorporated erano in corso studi su un calcolatore tascabile sperimentale a quattro funzioni basato su un circuito integrato singolo. Nel 1967 venne rilasciato un brevetto per questo dispositivo.

La storia della tecnologia dei semiconduttori è stata contrassegnata da incrementi regolari e perfino prevedibili nella complessità dei circuiti integrati, uniti a riduzioni di costo in pratica di tutti i componenti allo stato solido, inclusi i diodi a emissione di luce che, nella maggior parte dei calcolatori, presentano i numeri. Questo andamento evolutivo è stato spesso interpretato come il risultato di una «curva di apprendimento» determinata dall'esperienza acquisita nel corso di una produzione di volume continuamente crescente. In ogni modo verso il 1970 divenne possibile alloggiare l'intera logica di base del calcolatore su un solo chip di costo inferiore a 100 dollari e realizzato con la tecnologia dei semiconduttori a ossido-metallo. Nel giro di un anno questa tecnologia dava origine a una nuova generazione di piccoli calcolatori a quattro funzioni. Una macchina tipica aveva un circuito integrato singolo, che eseguiva tutte le funzioni di calcolo con l'ausilio di alcuni circuiti secondari.

Dei nuovi calcolatori alcuni erano tascabili e alimentati a batteria mentre altri erano macchine compatte da tavolo, alimentate dalla rete a corrente alternata. La loro caratteristica più importante era il prezzo (notevolmente inferiore a 200 dollari), che poneva i dispositivi nella fascia superiore del mercato di massa. Oggigiorno l'industria è avanzata lungo la curva di apprendimento di circa cento milioni di unità e la gamma dei prezzi si estende decisamente al di sotto dei 20 dollari con il risultato che per milioni di persone l'aritmetica di tutti i giorni non sarà mai più la stessa.

I piccoli calcolatori stanno subendo un'evoluzione così rapida e con sofisticazioni tanto differenziate che nessuna spiegazione del loro funzionamento sarebbe valida per tutti i tipi. Ciò nonostante è possibile spiegare i loro principi di funzionamento prendendo come esempio un ipotetico calcolatore a quattro funzioni costruito intorno a un tipico chip reale. Il chip non è l'ultima parola nel campo dei dispositivi microelettronici, ma è abbastanza rappresentativo dei circuiti integrati che sono tuttora impiegati dall'industria elettronica. È un chip semiconduttore a ossido-metallo con integrazione su larga scala, il che significa che su un quadrato di cinque millimetri di lato sono realizzati migliaia di componenti attivi e passivi: transistori, resistenze e diodi. Esso è dotato di 28 terminali, secondo lo standard dell'industria ed è alimentato con tensioni di circa sette volt. Il suo funzionamento è sincronizzato da un segnale alla frequenza di 250 chilohertz, cioè 250 000 cicli al secondo fornito da un *clock* (orologio campione), sotto forma di impulsi a intervalli di tempo regolari. Poiché chip analoghi sono stati già descritti in dettaglio in questa rivista (si veda l'articolo *La tecnologia a metallo-ossido-semiconduttore* di William C. Hattinger in «Le Scienze» n. 64, dicembre 1973), non ci si soffermerà qui sulla loro struttura fisica.

Oltre al chip, il calcolatore ha diversi altri componenti, alcuni visibili e altri no. Fra i primi quello che più attira l'attenzione è la tastiera per inserire le cifre e le istruzioni. Altri componenti sono il sistema di presentazione, il circuito oscillatore che genera il segnale di sincronismo, un regolatore di tensione per l'alimentazione, una serie di batterie ricaricabili e una custodia in plastica.

I 28 terminali del chip possono sembrare troppo pochi per ricevere e trasmettere tutte le informazioni che il chip stesso può trattare. Il motivo per cui sono sufficienti può essere spiegato nel modo migliore prendendo anzitutto in considerazione il sistema di presentazione del calcolatore. Se si osserva con attenzione la cifra 8 quando è illuminata si vede che essa è costituita da sette segmenti: tre nella parte superiore, tre in quella inferiore e uno al centro. Qualunque altra cifra, dallo 0 al 9, come pure il segno meno e uno qualunque dei vari simboli occorrenti per indicare errore o eccesso di capacità, può essere realizzata con meno di sette segmenti (*si veda l'illustrazione alle pagine 66 e 67*). Ogni segmento è un diodo LED, o diodo a emissione di luce, di forma allungata, in pratica un piccolo frammento di materiale semiconduttore con due terminali. Un ottavo diodo forma il punto decimale (che nella simbologia anglosassone ha il significato della nostra virgola), che può essere illuminato alla destra di ogni cifra.

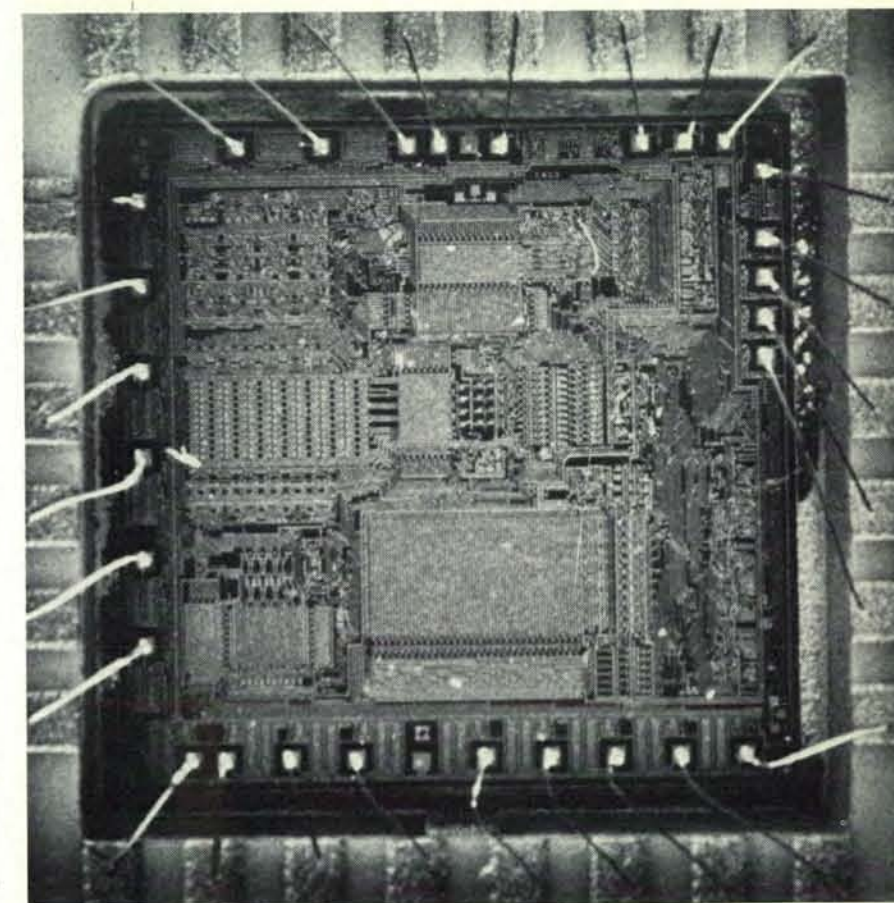
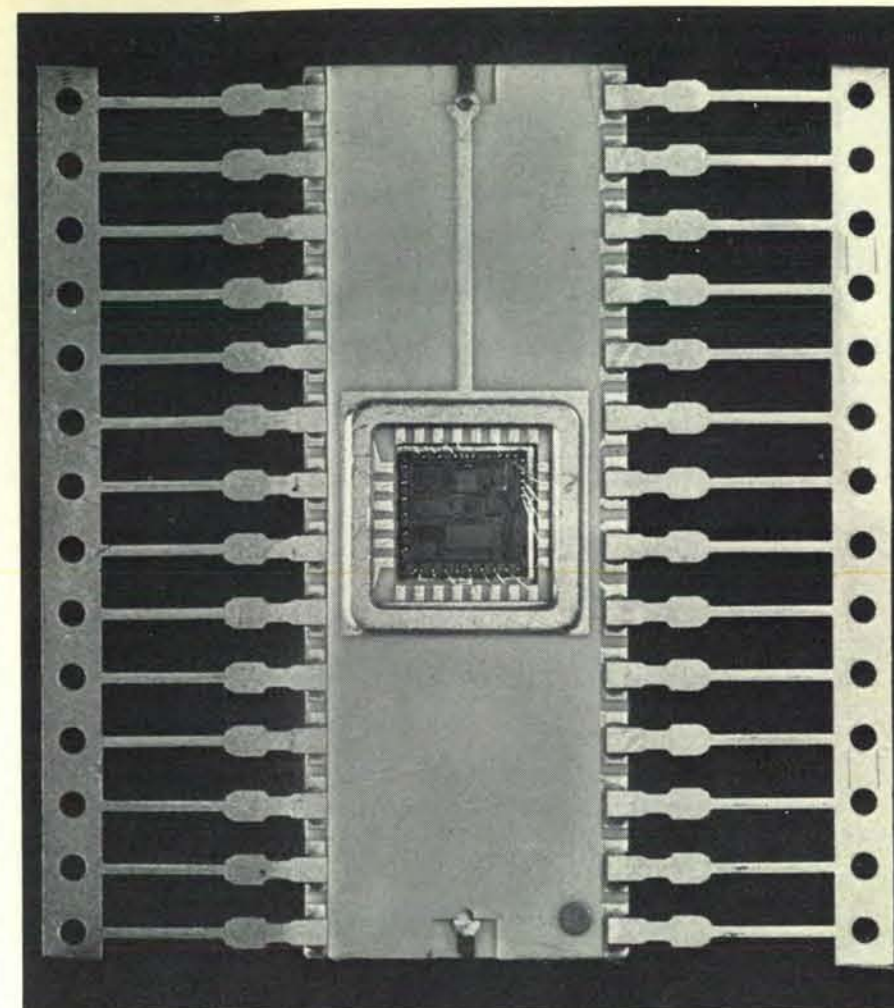
Il nostro ipotetico calcolatore possiede una fila di nove gruppi di otto diodi LED, per un totale di 72 diodi. I primi otto gruppi della fila, contando da destra, forniscono la presentazione delle otto cifre corrispondenti alla capacità del calcolatore, mentre il nono gruppo, l'ultimo a sinistra, serve per il segno meno, il punto decimale più a sinistra e i simboli di errore e di eccesso di capacità. Un diodo emette luce quando entrambi i suoi terminali sono alimentati, con il catodo a una tensione positiva rispetto all'anodo.

I 28 terminali del chip hanno anche la funzione di trasmettere gli ingressi dalla tastiera. Il calcolatore ha 18 interruttori a pulsante e 2 a cursore (oltre all'interruttore di accensione) ciascuno dotato di due terminali. La pressione esercitata su uno dei pulsanti o chiavi della tastiera chiude l'interruttore corrispondente.

Il problema di collegare fra loro 72 diodi LED e 20 interruttori tramite il chip e i suoi 28 terminali viene risolto nel modo seguente. La tastiera è collegata in effetti al chip tramite quattro terminali di ingresso solamente, indicati con *N, O, P*, e *Q*. I segmenti delle cifre costituiti dai diodi LED sono collegati al chip tramite otto terminali di uscita, indicati con lettere minuscole da *a* a *h*. A ogni istante il chip interagisce con non più di un interruttore per ogni linea di ingresso dalla tastiera e fa illuminare non più di uno dei nove gruppi di diodi. Ciò è reso possibile dall'esistenza di 11 linee di scansione, cioè di esplorazione, indicate con i numeri da 1 a 11: le linee di scansione hanno origine al chip collegandolo sia alla tastiera sia all'indicatore digitale.

Un dispositivo temporizzatore nel chip, sincronizzato dal segnale proveniente dal *clock* principale, alimenta una linea di scansione alla volta con un segnale impulsivo di tensione; la linea, normalmente alla tensione di massa, si trova durante l'impulso a una tensione positiva. Gli impulsi di tensione durano solo 132 microsecondi (pari a 33 cicli del *clock* principale). Un intervallo di sei cicli separa due successivi impulsi di scansione: un intero ciclo di scansione (durante il quale ogni linea di scansione viene alimentata una volta) dura quindi 429 cicli del *clock*, ovvero circa 1,7 millisecondi.

In quale modo le linee di scansione possono controllare 72 diodi LED tramite le otto uscite di segmento? Ciascuna delle nove linee di scansione da 1 a 8

Un chip microelettronico esegue le funzioni di calcolo di un calcolatore tascabile. La fotografia in alto mostra il chip quadrato, di cinque millimetri di lato, montato nella sua capsula di ceramica, sezionata per mostrare il chip. Il chip è collegato tramite 28 terminali con le altre parti del calcolatore. La microfotografia in basso mostra il chip ingrandito 17 volte. Il chip con la capsula, costruito dalla Texas Instruments Incorporated, costituisce un circuito integrato avente migliaia di funzioni elettroniche. Fra gli elementi strutturali del chip si riconosce, nell'area rettangolare in basso al centro, la memoria a sola lettura.

oltre alla linea 11 fornisce un ingresso a un circuito amplificatore e invertitore di corrente, esterno al chip principale, denominato pilota di cifra. Ciascun pilota, quando attivato da un impulso di scansione, fornisce una via verso massa, capace di una elevata densità di corrente, a tutti gli otto anodi di un gruppo di diodi LED. Pertanto la cifra che si trova in una certa posizione può essere illuminata per ogni ciclo di scansione solo durante il tempo in cui è attivata la corrispondente linea di scansione.

La particolare cifra presentata in un certo istante in una certa posizione è determinata da quali degli otto terminali di uscita di segmento risultano attivati durante un impulso di scansione. Ogni uscita di segmento controlla un circuito esterno amplificatore di corrente detto pilota di segmento. Ogni pilota di segmento è connesso al catodo dei nove diodi LED montati nella stessa posizione in ciascuno dei nove gruppi indicatori. L'attivazione di un pilota di segmento collega i nove catodi corrispondenti al polo positivo dell'alimentatore. La corrente scorre allora da questo particolare pilota di segmento attraverso l'unico diodo LED il cui anodo è collegato a massa tramite un pilota di cifra.

Un esempio chiarirà il funzionamento di questa disposizione: si supponga che siano attivi la linea di scansione *2* e le uscite *abcdgh*. Sull'indicatore compare (per 132 microsecondi) la cifra «3» nella seconda posizione da destra. La linea di scansione e l'indicazione vengono disattivati mentre nei 24 microsecondi successivi cambiano le uscite di segmento da attivare per indicare la cifra seguente a destra. Si supponga che la nuova configurazione sia *abcdef:* sull'indicatore apparirà nella prima posizione a destra «0», per 132 microsecondi (mentre la linea di scansione *1* è attiva). Malgrado che ciascuna delle cifre sia illuminata per circa l'otto per cento del tempo, la frequenza di accensione è così elevata che l'occhio vede indicato nelle ultime due posizioni a destra un «3.0» senza sfarfallamento.

Vediamo ora in quale modo le linee di scansione possono trasmettere i segnali provenienti dai 20 interruttori della tastiera agli appena quattro terminali di ingresso corrispondenti del chip. Ogni interruttore fornisce l'unica connessione possibile fra una data linea di scansione e una data linea di ingresso. A ogni istante i segnali di ingresso possono essere ricevuti solo dagli interruttori collegati alla linea di scansione che è attivata nello stesso istante (per la linea *1* si tratta degli interruttori catena/costante, del selettore decimale, della chiave «moltiplicare» e della chiave «1»). Un interruttore di questo gruppo che risulti chiuso quando viene attivata la particolare linea di scansione trasmette l'impulso alla corrispondente linea di ingresso.

Il chip principale, alla ricezione di un impulso su una linea di ingresso dalla tastiera, rileva quale linea di scansione è attiva in quell'istante e quindi decifra quali dei vari interruttori che potrebbero essere chiusi sulla linea di ingresso stessa sono effettivamente chiusi. Così durante un intero ciclo di scansione il calcolatore «giudica» ogni interruttore una volta. In questo modo i segnali provenienti dai 20 interruttori della tastiera vengono inoltrati con la tecnica della divisione di tempo ad appena quattro ingressi del chip principale.

Anche quando il calcolatore sta solamente presentando un numero, in attesa del successivo comando, il chip principale è attivo e segue la propria «procedura passiva» in fase con gli impulsi del *clock* principale. Durante tale procedura il chip esplora con continuità la tastiera e l'indicatore mentre le uscite di segmento vengono attivate e disattivate secondo ritmi precisi che originano una presentazione apparentemente stabile.

Oltre a mantenere la presentazione durante la procedura passiva, il chip deve elaborare i segnali di ingresso provenienti dalla tastiera. Durante la procedura passiva è in attesa di un segnale o sull'ingresso *N* (che lo connette a tutti i pulsanti o chiavi di cifra) o sull'ingresso *O* (connesso a tutti i pulsanti o chiavi di operazione). Quando su una di queste linee si presenta un impulso, il chip deve innanzitutto controllare che il segnale non sia soltanto un «rumore» casuale: il calcolatore verifica tale circostanza non tenendo conto del segnale ricevuto fin quando non ha verificato che il segnale è ancora presente alla fine del periodo di scansione successivo. Inoltre se una linea di scansione indica che sono chiusi e una chiave di cifra e una chiave di operazione, il chip tiene conto sempre del solo segnale di operazione.

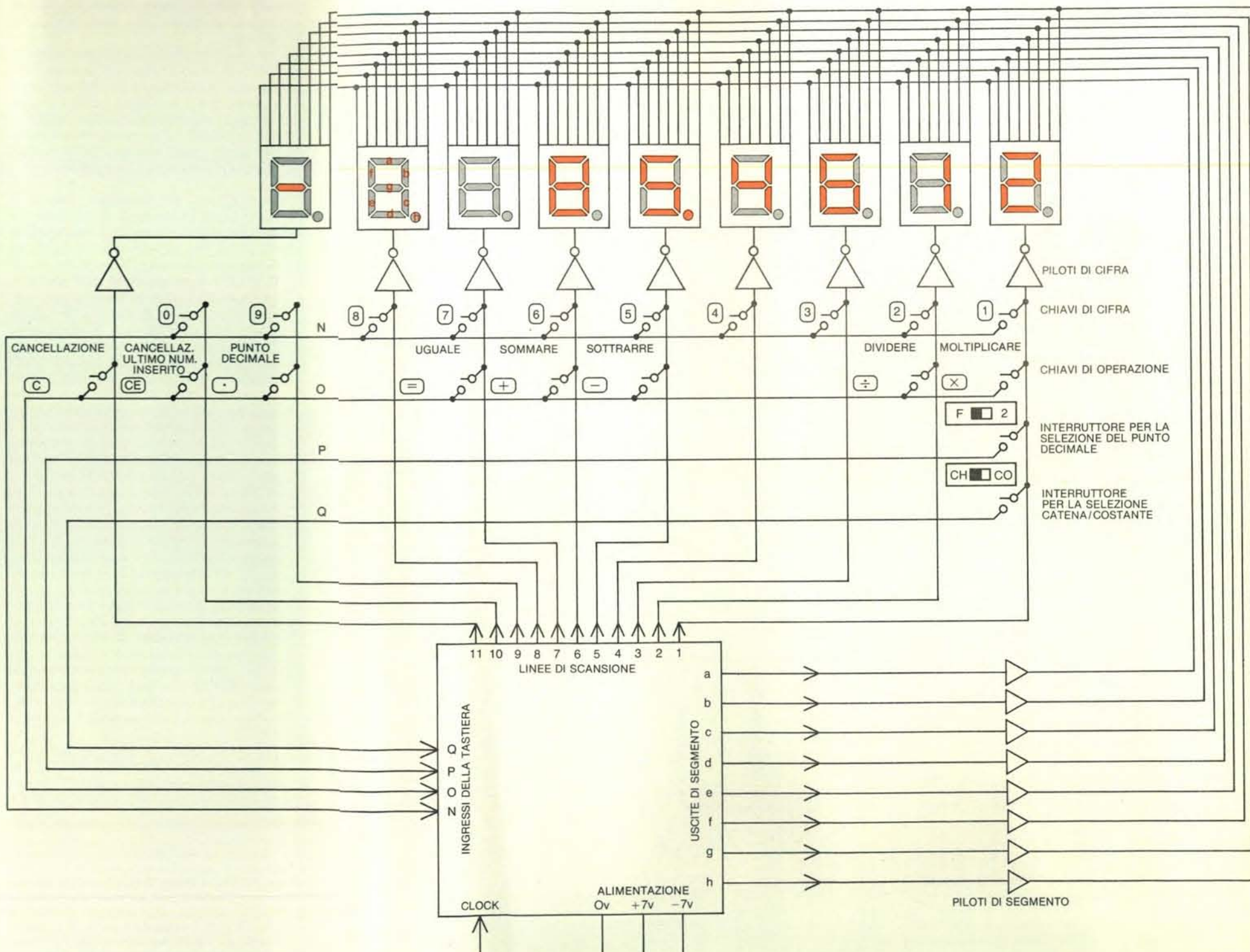
Mentre il chip elabora un segnale di cifra o un segnale di operazione, ignora ulteriori ingressi eventualmente presenti sulle linee *N* e *O*, ma continua sempre a controllare le linee *P* e *Q* per rilevare la presenza di istruzioni relative alla posizione del punto decimale e al tipo di calcolo (catena o costante).

Un chip può completare un'operazione anche prima che l'operatore rilasci il pulsante che ha avviato l'operazione medesima. Ciò significa che lo stesso comando non deve essere azionato più di una volta. Perciò il chip, prima di riprendere la procedura passiva, verifica che tutti i pulsanti siano stati rilasciati.

Da quanto precede risulta evidente che il calcolatore compie un numero di operazioni ben superiore alle semplici addizione, sottrazione, moltiplicazione e divisione, essendo in realtà un sistema di processo dell'informazione multiruolo di notevole sofisticazione. La chiusura degli interruttori e la presentazione sono operazioni importanti, ma ciò che consente all'ipotetico calcolatore di funzionare sono l'architettura, o organizzazione logica, e gli algoritmi o procedure numeriche. Nell'illustrazione a pagina 71 sono schematizzati i sottosistemi elettronici più importanti facenti parte del chip principale. Nel seguito verranno descritti i compiti della maggior parte dei sottosistemi e si accennerà al modo in cui questi eseguono un semplice calcolo.

Il cuore del sistema, per quanto concerne il funzionamento, è l'addizionatore-sottrattore, chiamato in genere semplicemente addizionatore, che opera su numeri espressi in codice binario. Nel codice numerico binario puro ogni bit, o cifra binaria, rappresenta una potenza di 2. Un numero binario, per esempio 10010, si legge nel modo più semplice da destra verso sinistra: il primo bit indica la presenza di 1, il secondo la presenza di 2, il terzo la presenza di 4 e così via. La notazione binaria 10010 quindi rappresenta la codificazione di un numero in codice decimale composto da nessun 1, un 2, nessun 4, nessun 8 e un 16: in altre parole il numero decimale è 0 + 2 + 0 + 0 + 16 cioè 18.

Il nostro ipotetico calcolatore impiega una variante di questo sistema conosciuto come notazione BCD (binary-coded-decimal, cioè decimale codificata binaria). Nel sistema BCD ogni cifra di un numero decimale è rappresentata da un codice binario puro di quattro bit. Quindi il numero 97 in codice BCD è 1001 0111. La cifra 1 è rappresentata da una tensione più alta e la cifra 0 da una tensione più bassa; nel sistema non esistono tensioni intermedie.

Quando deve essere eseguita una operazione di somma o di differenza, all'addizionatore vengono presentati due numeri interi in codice BCD sotto forma di una coppia di cifre a quattro bit per volta, cominciando dalle cifre meno significative. Per esempio, se si vuole sommare 56 a 43, che sono espressi in codice BCD rispettivamente da 0101 0110 e 0100 0011, le cifre a destra (ossia 0110 e 0011) vengono presentate per prime all'addizionatore.

Nello schema sono riportati i circuiti della tastiera e dell'indicatore esterni al chip principale, rappresentato dal quadrato in basso al centro. Tramite questi circuiti, il chip comunica con 20 interruttori della tastiera e con 72 diodi a emissione di luce (LED) tramite solo 23 dei suoi 28 terminali. Nella matrice formata dalle 11 linee di scansione e dalle quattro linee di ingresso dalla tastiera, ciascuno dei 44 incroci è una possibile ubicazione per un interruttore. Per questo calcolatore sono necessari solo 20 interruttori. Quando è in funzione, le linee di scansione sono alimentate ciclicamente, una alla volta, per un intervallo di 132 microsecondi, cosicché le cifre illuminate nell'indicatore vengono presentate con un impercettibile sfarfallamento. Una cifra è realizzata con un massimo di 7 diodi LED, che sono tutti illuminati quando viene presentata la cifra otto. Un ottavo diodo serve per il punto decimale.

SIGNIFICATO DEI NUMERI	POSTI NEL REGISTRO				FASI DURANTE L'ADDIZIONE IN UN PASSO NELL'ADDIZIONATORE
	MIGLIAIA	CENTINAIA	DECINE	UNITA'	
PRIMO ADDENDO 0835	0000	1000	0011	0101	NUMERI IMMAGAZZINATI NEI REGISTRI DI OPERANDO E DI INDICATORE CHE DEVONO ESSERE SOMMATI IN UN PASSO NELL'ADDIZIONATORE
SECONDO ADDENDO 0974	0000	1001	0111	0100	
5				0101	SOMMA DELLE CIFRE MENO SIGNIFICATIVE
4				0100	
SOMMA 9, RIPORTO 0			0	1001	LA SOMMA BINARIA NON È MAGGIORE DI 9 E QUINDI VIENE TRASFERITA NELL'ACCUMULATORE MENTRE IL RIPORTO VIENE IMMAGAZZINATO IN UN LATCH INTERNO
3			0011		SOMMA DELLE DUE CIFRE SUCCESSIVE E DEL RIPORTO
7			0111		
BINARIO PURO 10		0	1010		LA SOMMA BINARIA È MAGGIORE DI 9 E QUINDI VIENE AGGIUNTA LA CORREZIONE 6 IN BCD
CORREZIONE 6		0	0110		
SOMMA 0, RIPORTO 1		1	0000		IL RIPORTO VIENE IMMAGAZZINATO NEL LATCH E LA SOMMA NEL REGISTRO DELL'ACCUMULATORE
8		1000			SOMMA DELLE DUE CIFRE PIÙ SIGNIFICATIVE E DEL RIPORTO
9		1001			
BINARIO PURO 18	1	0010			LA SOMMA BINARIA È MAGGIORE DI 9 E QUINDI VIENE AGGIUNTA LA CORREZIONE 6
CORREZIONE 6	0	0110			
SOMMA 8, RIPORTO 1	1	1000			LA SOMMA FINALE E IL RIPORTO VENGONO TRASFERITI NEL REGISTRO DELL'ACCUMULATORE
SOMMA 1809	0001	1000	0000	1001	SOMMA FINALE NEL REGISTRO DELL'ACCUMULATORE

Un esempio di addizione mostra come i numeri in codice decimale 835 e 974 vengono addizionati da un calcolatore. Il calcolatore opera su numeri in codice BCD (binary coded decimal) nei quali ogni cifra del numero decimale è espressa con quattro cifre binarie. Per esempio il 5 di 835 è espresso come 0101, che letto da destra significa un 1, nessun 2 e un 4. Ogni gruppo di quattro bit è tuttavia sommato necessariamente in codice binario puro. Così 3 (0011) più 7 (0111) dà luogo a 1010, che esprime il numero 10 in codice binario puro, ma che non ha significato in BCD. Pertanto alla rete addizionatrice BCD viene fornito un secondo addendo per sommare 6 (0110) a qualsiasi somma maggiore di 9 espressa in binario puro per convertirla in codice BCD. In questo caso il numero espresso in binario puro 10 (1010), una volta sommatogli 6 (0110), dà il binario puro 16 (10000), che è interpretato correttamente in codice BCD come 10000, ovvero 10. Due correzioni BCD di questo tipo sono necessarie in questo esempio mentre il calcolatore tascabile esegue l'operazione di sommare 835 e 974.

registro può scorrere in un senso o in quello opposto rispetto a quelli contenuti negli altri registri. Infine un ingresso ai circuiti di avviamento è riservato a una cifra sintetizzata da altri circuiti del blocco di controllo: si tratta del percorso lungo il quale una cifra decodificata da un ingresso dalla tastiera può essere avviata attraverso il sistema dei registri.

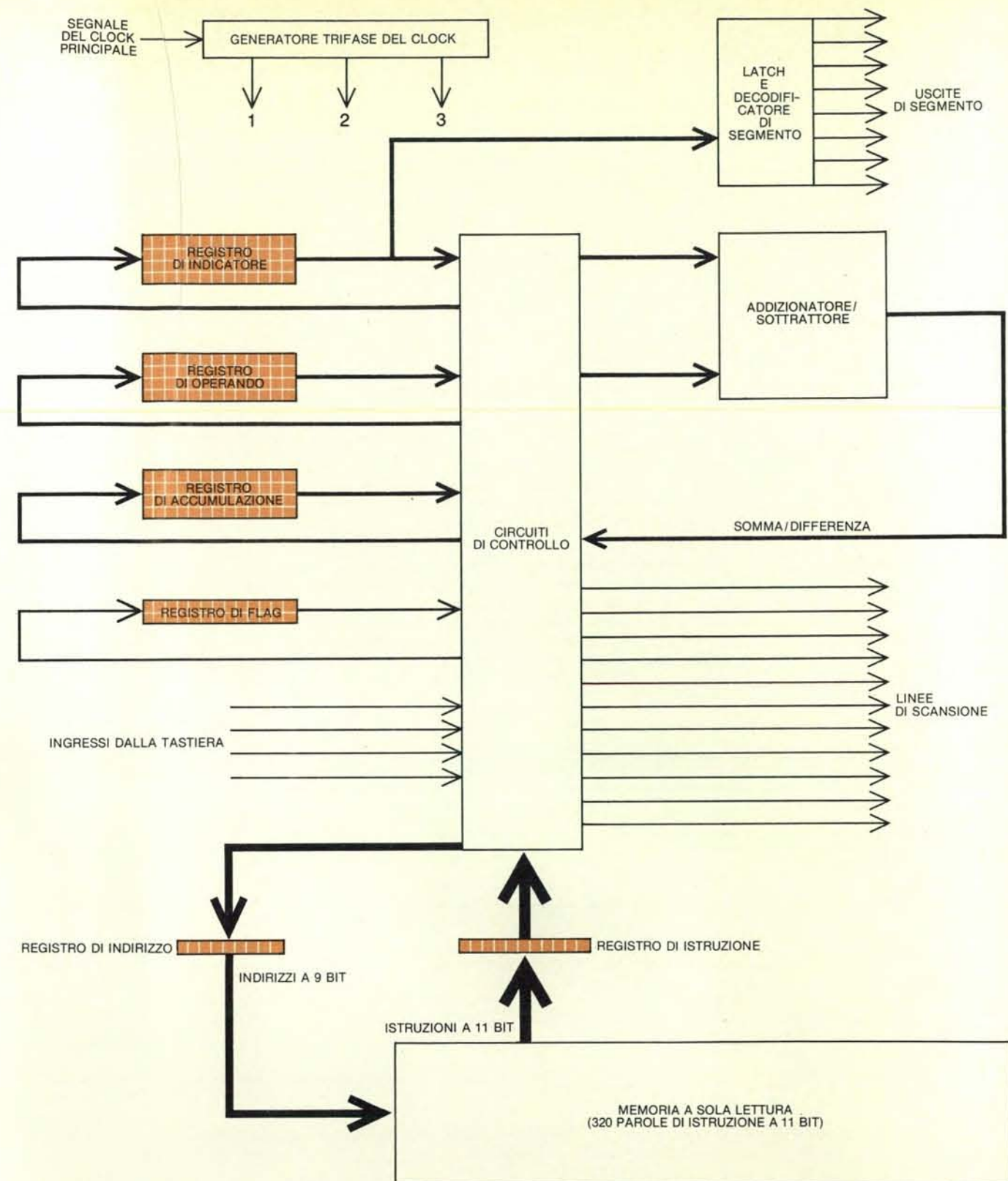
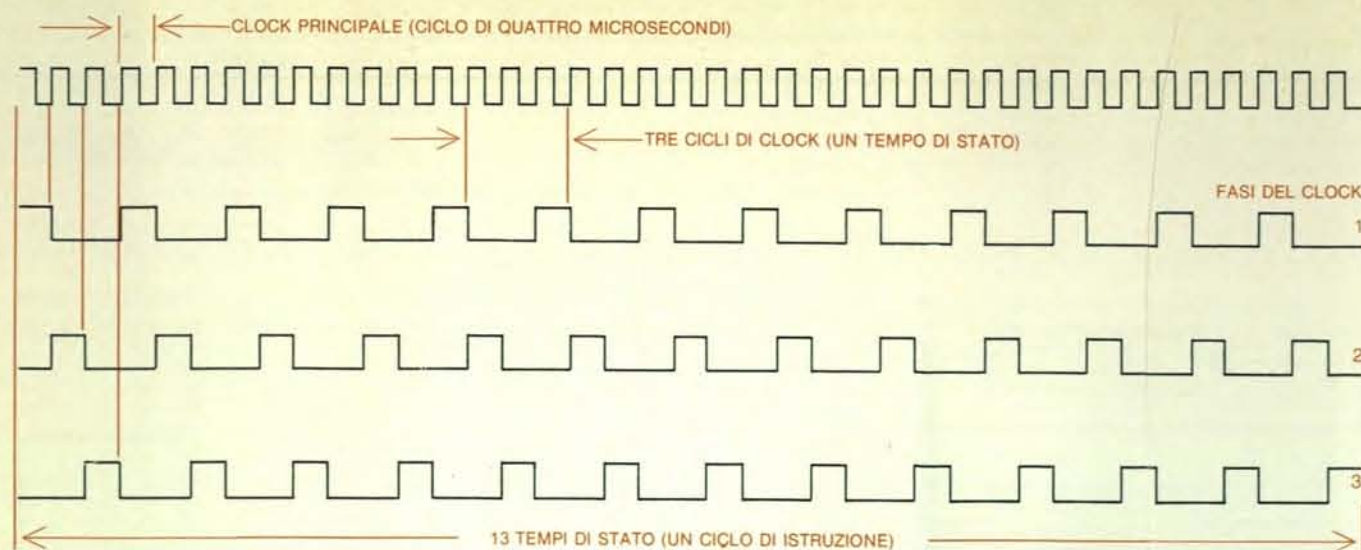
I circuiti di decodificazione, di temporizzazione e di controllo, che possono essere chiamati nel loro insieme il regolatore, regolano il funzionamento di tutti gli altri sottosistemi del chip. Le sue funzioni sono governate essenzialmente da una o l'altra delle 320 parole di istruzione (ciascuna costituita da 11 bit) fornite al registro di istruzione (e da questo al regolatore) dalla memoria a sola lettura (ROM, *read-only memory*), così chiamata perché contiene una serie programmata di istruzioni che non può essere modificata dopo la costruzione del calcolatore. Ogni parola di istruzione, ottenuta dalla memoria a sola lettura mediante parole di indirizzo a nove bit, stabilisce le regole di funzionamento valide durante un ciclo di istruzione della durata di 13 tempi di stato (39 cicli del *clock*).

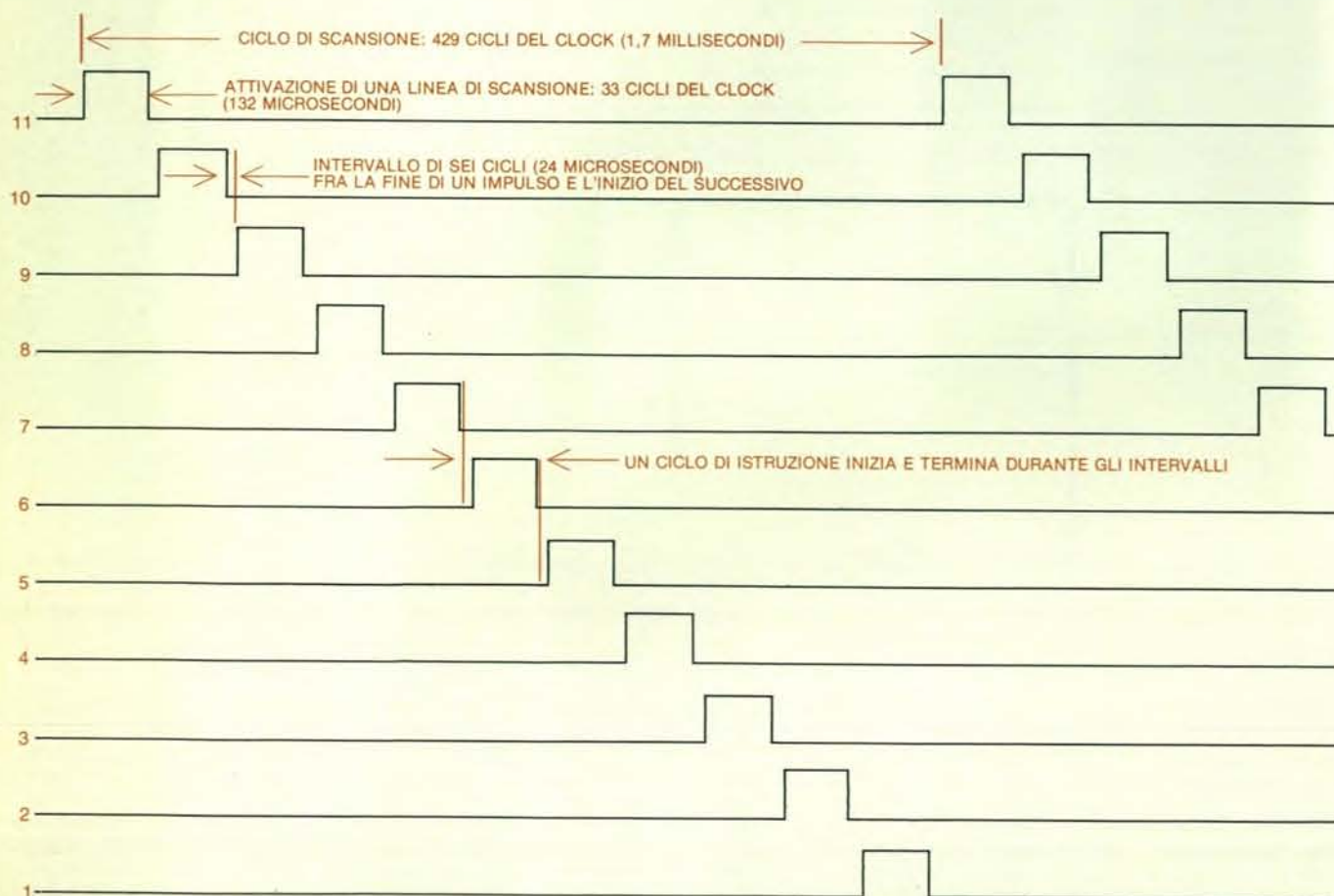
Durante ogni ciclo di istruzione il registro a scorrimento può effettuare non solo una circolazione completa, ma anche una addizione o una sottrazione, se è richiesta un'operazione di uno di questi tipi. Durante il ciclo si possono avere anche scorrimenti e scambi dei registri. Altri particolari delle azioni eseguite durante un ciclo di istruzione sono condizionati dai segnali che giungono al chip attraverso le quattro linee provenienti dalla tastiera, dal conteggio degli impulsi di *clock*, dal bit di riporto o di prestito proveniente dall'ultima operazione dell'addizionatore e dalle 13 successive coppie di bit contenute in un registro a scorrimento a 2 per 13 bit che serve come una specie di memoria «tampone» a disposizione del regolatore. Quest'ultimo è il componente noto come registro di *flag*.

Si supponga di richiedere al calcolatore di eseguire il calcolo 2 + 3, che implica, per i circuiti del chip, lo stesso tipo di procedura che sarebbe seguita in un calcolo molto più complicato. Dopo avere acceso la macchina si preme il pulsante «2». Il segnale così attivato viene inviato tramite una delle linee di ingresso dalla tastiera al regolatore, dove determina l'inserimento di un 2 nel registro di indicatore. Da questo i numeri (e quindi anche il 2) sono fatti pervenire all'indicatore digitale luminoso del calcolatore.

Si preme poi il pulsante «più»: il segnale che viene attivato è immagazzinato come un codice nel registro di *flag*, a disposizione del regolatore per consultazione. Successivamente si preme il pulsante «3». Il 2 già immagazzinato nel registro di indicatore viene trasferito nel registro di operando, mentre il 3 segue le stesse fasi già attraversate in precedenza dal 2 per terminare nel registro di indicatore ed essere da qui trasmesso all'indicatore luminoso.

Si preme infine il pulsante «uguale»: il regolatore consulta il registro di *flag* per stabilire quale operazione è stata richiesta sui due numeri (2 e 3) che stanno ora circolando nei registri di indicatore e di operando e si accerta che debbano essere addizionati. Il regolatore estrae i numeri dai registri e li presenta all'addizionatore, il 2 come 0010 e il 3 come 0011. Qui vengono sommati e il risultato è inviato al registro di accumulazione, e da questo traslato al registro di indicatore e all'indicatore luminoso.

Calcoli anche di poco più complessi di 2 + 3 possono richiedere numerosi passi attraverso i circuiti di avviamento e di ritardo ed eventualmente attraverso

In questo schema è rappresentata l'organizzazione concettuale del chip. Ogni blocco rappresenta uno dei sottosistemi elettronici principali. Lo spessore delle frecce indica la quantità di informazioni che viaggia simultaneamente da una sezione all'altra; le frecce più sottili rappresentano un bit, o cifra binaria, e quelle più spesse, che escono verso l'alto dalla memoria a sola lettura (ROM) rappresentano il bit in parallelo. Per esempio, per sommare 5 + 4, l'operatore comincia premendo il pulsante «5» della tastiera. Il segnale viene inviato tramite una delle linee di ingresso dalla tastiera al circuito di controllo, dove (secondo le opportune istruzioni provenienti dalla memoria a sola lettura) viene codificato in forma binaria e trasmesso al registro di indicatore. L'operatore preme poi il pulsante «più» e l'informazione viene immagazzinata nel registro di *flag* (o di segnalamento). Quando viene premuto il pulsante «4» il segnale viene manipolato in modo analogo al precedente segnale proveniente dal pulsante 5, mentre il 5 immagazzinato nel registro di indicatore viene trasferito nel registro di operando. Infine l'operatore preme il pulsante «uguale». Il circuito di controllo, sempre seguendo le istruzioni della memoria fornite fase per fase, si accerta dal registro di *flag* dell'operazione che deve essere compiuta sui due numeri immagazzinati, li presenta all'addizionatore che a sua volta li somma e invia quindi il risultato dell'operazione nel registro di accumulazione. Da qui il risultato viene trasmesso immediatamente al registro di indicatore, e da questo al *latch* e decodificatore di segmento per poter essere infine presentato sul pannello indicatore.

Il clock principale di un tipico calcolatore a quattro funzioni opera alla frequenza di 250 chilohertz, ovvero 250 000 cicli al secondo. Fra le altre attribuzioni, il clock stabilisce la temporizzazione per la circolazione dei dati fra i registri di indicatore, di operando, di accumulazione e di *flag*, che costituiscono in realtà un unico registro a 14 per 13 bit (la memoria ad accesso casuale) che compie con continuità uno scorrimento di un passo ogni tre impulsi del clock principale. Tre di questi impulsi costituiscono un «tempo di stato» e 13 tempi di stato consecutivi (uno per ogni cifra immagazzinata in un registro a 13 bit) formano quello che viene solitamente detto un ciclo di istruzione.

l'addizionatore. Si supponga che il primo numero del calcolo sia 25,6. Viene anzitutto inserito il 2, che appare sull'indicatore con un punto decimale alla sua destra. Quando si inserisce il 5, il 2 deve scorrere di un posto verso sinistra, ma il punto decimale deve rimanere nella posizione precedente. Viene poi premuto il tasto del punto decimale; l'indicatore non cambia la presentazione, ma il segnale proveniente dal pulsante viene immagazzinato come un segno nel registro di *flag*. Ora quando si preme il pulsante «6», il punto decimale deve scorrere di un posto verso sinistra, insieme con il 2 e il 5. Lo scorrimento del punto decimale richiede che il calcolatore esegua un ciclo di addizione, applicato al simbolo del punto decimale che delle 13 cifre nel registro di indicatore è quella impiegata per seguire la posizione del punto.

Il ciclo di scansione occupa, in un calcolatore tipico, 429 cicli del *clock*, di quattro microsecondi ciascuno. Ogni traccia indica la tensione in una linea di scansione in funzione del tempo e gli impulsi verticali mostrano la successione temporale degli impulsi di attivazione.

Se il successivo comando dalla tastiera è di eseguire una addizione e il numero da sommare a 25,6 è 33,14, il primo compito è di allineare i punti decimali nei due registri interessati, nello stesso modo in cui si opererebbe eseguendo un'operazione manuale. In altre parole le cifre del punto decimale nei due registri i cui contenuti sono da sommare devono trovarsi nella stessa posizione: perciò il numero 25,6 deve essere scorso di un posto verso sinistra, aggiungendo una cifra ai decimali (con risultato 25,60.). Compiuta questa fase, i due numeri possono venire addizionati in un solo passo tramite l'addizionatore.

Gli esempi precedenti dovrebbero essere sufficienti per indicare il tipo di operazioni, verifiche e collegamenti sequenziali che devono essere eseguiti in risposta a un ingresso dalla tastiera. Una addizione o sottrazione completa con un punto decimale fluttuante può richiedere perfino 300 cicli di istruzione (equivalenti a 12 000 cicli del *clock*): il numero esatto dipende dal tipo di programmazione e dalla dimensione dei numeri su cui deve essere eseguito il calcolo. La moltiplicazione e la divisione, date da addizioni, sottrazioni e scorrimenti iterativi, richiedono un numero proporzionalmente maggiore di cicli di istruzione.

Sebbene 12 000 cicli del *clock* possano sembrare un tempo piuttosto lungo per eseguire una semplice addizione o sottrazione, si deve ricordare che il tempo effettivamente richiesto è minore di 0,05 secondi. In verità si tratta di un tempo che potrebbe sembrare interminabile per un grande calcolatore elettronico, con una frequenza di *clock* all'incirca 100 volte più alta di quella del calcolatore in questione e dotato di una programmazione più efficiente, soprattutto per l'uso di parole di istruzione più lunghe. All'operatore umano sembra che il calcolatore esegua le operazioni istantaneamente.

Il nostro ipotetico calcolatore, come la maggior parte degli altri calcolatori, presenta molte caratteristiche di progettazione che derivano dai grandi calcolatori. Per esempio l'addizionatore, insieme con le parti del regolatore che instradano e ritardano le cifre in transito, corrisponde all'unità centrale (CPU - *Central Processing Unit*), e i tre registri dei dati alla memoria principale. Le altre parti del regolatore sono equivalenti ai circuiti temporizzatori e di controllo che guidano il funzionamento del calcolatore al livello più basso, cioè al livello in corrispondenza del quale il cambiamento delle operazioni richiede modifiche nella configurazione del cablaggio. Le procedure programmate immagazzinate nella memoria a sola lettura sono grosso modo equivalenti ai microprogrammi di certi calcolatori (un microprogramma di un calcolatore è una procedura più o meno fissa, che governa direttamente i circuiti di controllo e che viene chiamata in causa da un'istruzione di un programma espresso in linguaggio di macchina).

L'esecuzione di un microprogramma in un calcolatore assomiglia a un programma vero poiché in generale richiede di cercare un'istruzione dopo l'altra in una memoria di un qualche tipo in conformità con indirizzi definiti in parte dal controllo dei risultati di operazioni precedenti. Tuttavia il concetto di programma racchiude in generale l'idea di una serie di istruzioni che possono essere ordinate e alterate all'esterno della macchina stessa più facilmente di quanto non avvenga per i microprogrammi. La distinzione è evidente nei calcolatori «programmabili», nei quali ogni istruzione di un programma (possibilmente una composta da un operatore) dà luogo tipicamente a una procedura microprogrammata del tipo descritto. La distinzione diviene alquanto vaga quando si prende in esame il modo in cui i calcolatori elettronici eseguono operazioni diverse dalle quattro fondamentali, come estrarre una radice quadrata, calcolare un seno o y^x. Operazioni di questo tipo sono in generale effettuate seguendo una sequenza programmata (realizzata con microprogrammi) di addizioni, sottrazioni, moltiplicazioni e divisioni.

Un risultato della analogia di struttura tra i grandi e i piccoli calcolatori elettronici è che questi ultimi, come quelli di maggiori dimensioni, possono eseguire una varietà di procedure programmate differenti usando la stessa struttura di circuiti elettronici. In un grande calcolatore i differenti programmi sono introdotti con istruzioni che vengono immagazzinate nella memoria, mentre i microprogrammi di un piccolo calcolatore sono immessi nella sua memoria a sola lettura durante la fabbricazione.

Si possono distinguere diverse linee nell'evoluzione del piccolo calcolatore elettronico, tutte convergenti nel disporre di un maggior numero di funzioni e in una maggiore utilità a costi minori. Le fonti di questi progressi sono le riduzioni di dimensioni, l'esperienza acquisita nella produzione di grande serie e i progressi della tecnologia dei semiconduttori.

Da una parte si sono realizzate notevoli riduzioni di costo con provvedimenti in apparenza insignificanti, come il disporre le piste dei circuiti stampati perfettamente ortogonali in modo che non esistano due linee di scansione o due linee di ingresso dalla tastiera che si incrocino. In altre parole uno schema della matrice della tastiera corrisponde esattamente alla configurazione effettiva della tastiera stessa. La soluzione rende più complessa la programmazione delle procedure avviate dagli ingressi dalla tastiera, ma riduce il numero di strati sovrapposti necessari per la piastra circuito stampato sotto le chiavi, tecnica questa di costo rapidamente crescente con l'aumento del numero degli strati. D'altra parte innovazioni tecnologiche di vasta portata possono rendere ben presto pratica, da un punto di vista economico, l'integrazione di tutti i circuiti elettronici del calcolatore in un solo chip, con conseguente riduzione nel costo di montaggio e aumento dell'affidabilità.

Le variazioni cromosomiche nell'evoluzione dei primati

Lo studio delle modalità di variazione dei caratteri ereditari nei primati non-umani permette di formulare alcune ipotesi sugli eventi che hanno condotto alla differenziazione dell'uomo

di Brunetto Chiarelli

Il gruppo zoologico dei primati presenta, dal punto di vista genetico, un notevole interesse; infatti le forme attualmente viventi rispecchiano con una certa approssimazione alcune delle tappe che hanno portato all'evoluzione della nostra specie.

La sistematica delle forme viventi dei primati riproduce cioè, nei caratteri essenziali, quella delle forme fossili e questo spiega l'interesse particolare che questo gruppo offre allo zoologo.

Lo studio comparativo dei caratteri ereditari, comuni alle diverse specie di scimmie, ci sembra interessante, sia perché permette di ricostruire il tracciato seguito nell'evoluzione dalla comparsa alla stabilizzazione di ogni carattere, sia perché le scimmie sono animali più vicini all'uomo di quanto non lo siano la drosofila o il topo.

Il genetista umano in altre parole, potrà meglio apprezzare i limiti e le modalità di variazione dei caratteri ereditari conoscendo le modalità di variazione nelle specie a noi più vicine. Naturalmente la presenza di un gene in due specie differenti non implica necessariamente una recente separazione filogenetica: porzioni omologhe di genoma si possono trovare in specie anche lontane. D'altra parte, specie affini o popolazioni appartenenti alla medesima specie possono talvolta differire per la perdita o la modificazione di uno o più geni.

La genetica comparata dei primati è tuttavia una scienza recente e i dati di cui dispone non sono ancora ben coordinati. È ancora prematuro pensare di trovare dei gruppi d'associazione, e ancora di più pretendere di stabilire le mappe genetiche delle varie specie di scimmie quando queste a malapena si cominciano a conoscere per l'uomo. Tuttavia i primi risultati sono promettenti. Infatti lo studio comparato dei prodotti di un gene ci permette già di scoprire con più facilità le singole mutazioni che conducono alla differenziazione tra gli individui e i meccanismi che tendono a separare le popolazioni e a formare nuove specie.

In effetti una singola mutazione genica in un organismo superiore, raramente crea una «barriera riproduttiva» sufficiente perché si origini una nuova specie. Inoltre le mutazioni sono rare e sono soggette a una forte pressione selettiva.

La genetica comparata dei primati non umani ci interessa per i contributi che esso può offrire alla genetica umana e alle sue applicazioni cliniche, ma soprattutto per chiarire le relazioni filogenetiche intercorrenti tra le diverse specie. Questo è particolarmente evidente nello studio dei cromosomi che sono i veicoli dell'informazione ereditaria.

Ciascuna cellula del nostro organismo, come quelle di tutti gli altri esseri viventi, contiene un corpuscolo che si può colorare con coloranti basofili ed è costituito prevalentemente da acido desossiribonucleico (DNA): il nucleo. Esso appare pressoché omogeneo in tutte le cellule non in divisione, ma, quando la cellula si divide, all'interno del nucleo si formano delle strutture filamentose: i cromosomi. Quest'ultimi caratterizzano la specie a livello cellulare. Ciascuna specie animale o vegetale è caratterizzata da un determinato numero di cromosomi e da una loro morfologia tipica.

Sino a circa 15 anni fa, lo studio dei cromosomi dei mammiferi era molto difficile, ma in seguito, con l'introduzione delle colture *in vitro* e del trattamento ipotonico, queste ricerche hanno progredito in modo notevole. Fu nel 1956 che Tijo e Levan descrissero per primi il numero esatto dei cromosomi umani.

L'uomo possiede 46 cromosomi e il loro numero e la loro morfologia sono costanti a tal punto che la presenza supplementare di uno dei più piccoli (il ventunesimo) nelle cellule di un individuo, causa la sindrome di Down o mongolismo (come dimostrarono Lejeune e Turpin nel 1959). La loro funzione è di conseguenza di grande importanza nell'organizzazione generale della vita di un individuo e per l'insieme degli individui costituenti la specie. Le informazioni ereditarie sono localizzate in successione lineare su di essi.

Nel corso della metafase della divisione mitotica, i cromosomi sono particolarmente visibili e risultano costituiti da due filamenti (cromatidi) che si congiungono in una regione caratteristica di ciascun cromosoma (centromero).

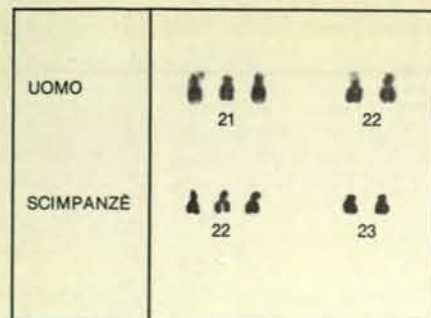
A questo stadio, ciascun cromosoma si è già autoduplicato e ciascuno dei cromatidi si trasferirà in una delle due cellule figlie. Con questo meccanismo di autoduplicazione il patrimonio di informazioni che l'individuo riceve dai suoi genitori, al momento del concepimento, rimane inalterato nel corso delle successive generazioni cellulari per tutta la durata della sua vita.

Perché la quantità di informazioni trasmessa da un individuo all'altro rimanga rigorosamente costante, è necessario che il numero dei cromosomi, durante la maturazione delle cellule germinali (spermatozoi e uova), si riduca a metà, di modo che si possa nuovamente ricostituire nello zigote un patrimonio genetico identico al precedente.

A differenza delle cellule somatiche, le cellule germinali hanno la proprietà di trasmettere perpetuamente i loro cromosomi da un individuo all'altro nel corso delle varie generazioni. Le cellule somatiche di un individuo sono destinate a morire con la sua morte, al contrario le cellule germinali sono potenzialmente immortali.

Sono dunque queste ultime che stabiliscono una continuità da un individuo

I primati rappresentano un gruppo zoologico particolarmente interessante: la sistematica delle forme attualmente viventi ripete pressappoco i vari stadi evolutivi attraverso i quali questo gruppo è passato durante i suoi 80 milioni di anni di esistenza. Le proscimmie in qualche modo rappresentano un gruppo di forme più primitive delle scimmie; le scimmie platirrine uno stadio attraverso il quale molte delle forme di scimmie catarrine sono dovute passare; queste ultime costituiscono uno stadio evolutivo delle antropomorfe e queste infine uno stadio della evoluzione dell'uomo. Una sequenza quindi di forme che, in qualche modo, ripete i numerosi stadi evolutivi che dal primitivo *Purgatorius* del Cretacico superiore hanno condotto all'attuale *Homo sapiens*.

	SOTTORDINI	FAMIGLIE	GENERI
ORDINE PRIMATI	PROSCIMMIE	*TUPAIIDAE*	*TUPAIA*
		LORISIDAE	*LORIS* (1)
		GALAGIDAE	*GALAGO* (2)
		LEMURIDAE	*LEMUR* (3)
		INDRIIDAE	*INDRI* (4)
		DAUBENTONIIDAE	*DAUBENTONIA*
		TARSIIDAE	*TARSIUS*
	SCIMMIE PLATIRRINE	*CALLITHRICIDAE*	*CALLITRIX* (5)
		CALLIMICONIDAE	*CALLIMICO*
		CEBIDAE	*CEBUS* (6)
	SCIMMIE CATARRINE	*CERCOPITHECIDAE*	*MACACA* (7)
			PAPIO
			THEROPITHECUS (8)
			CERCOCEBUS
			CERCOPITHECUS (9)
		COLOBIDAE	*PRESBYTIS* (10)
			PYGATRIX
			RHINOPITHECUS (11)
			SIMIAS
			NASALIS (12)
			COLOBUS
		HYLOBATIDAE	*HYLOBATES* (13)
			SYMPHALANGUS
			NOMASCUS
		PONGIDAE	*PONGO*
			PAN (14)
			GORILLA
		HOMINIDAE	*HOMO*

Confronto tra i cromosomi del gruppo G umani e di scimpanzè. Sia la coppia 21 dell'uomo che la coppia 22 dello scimpanzè in questo esempio presentano un piccolo cromosoma acrocentrico soprannumerario collegato a una grave anomalia fenotipica che, nell'uomo, prende il nome di sindrome di Down.

all'altro, nel corso delle varie generazioni e soltanto le variazioni che si attuano nei cromosomi delle cellule germinali possono produrre delle variazioni nel complesso ereditario di una specie.

Durante la maturazione delle cellule germinali, i cromosomi passano attraverso i vari stadi che conducono alla riduzione a metà del corredo cromosomico o, come si dice in gergo tecnico, da un assetto diploide (2n) a uno aploide (n).

Allorché uno spermatogonio cessa di moltiplicarsi per mitosi e comincia a ingrossare, significa che la divisione meiotica sta per iniziare. La cellula che si trova a questo stadio viene detta «spermatocita I», ed è destinata a formare 4 spermatozoi. Il risultato della divisione meiotica è la riduzione a metà del numero dei cromosomi.

Lo stato aploide che si stabilisce nello spermatozoo, costituisce l'apporto del maschio al patrimonio ereditario dei suoi figli. L'oogenesi differisce dalla spermatogenesi soprattutto per il comportamento del citoplasma durante la divisione meiotica. Per l'accumulo di sostanze nutritizie, l'oocita primario è decisamente più grande dello spermatocita primario. Il comportamento dei cromosomi durante la maturazione delle uova è tuttavia identico a quello degli spermatozoi. Poiché durante le due divisioni meiotiche ha luogo una sola divisione dei cromosomi, questa provoca un dimezzamento del numero dei cromosomi dei gameti rispetto alle altre cellule dell'organismo.

A un certo stadio della divisione, tuttavia, ognuno dei due omologhi si deve separare e ciascun gamete possiederà un solo elemento per ogni coppia.

È il caso che decide quale dei due cromosomi omologhi si trasferirà in un determinato gamete, di modo che le diverse possibilità (combinazioni) si realizzano con uguale frequenza.

Questo è un fattore che contribuisce notevolmente a mantenere l'associazione casuale dei caratteri ereditari all'interno di una specie. Ma vi è un altro meccanismo che tende a rendere ancora più casuale la combinazione dei geni nella discendenza. Questo meccanismo consiste nello scambio di porzioni di cromosoma tra gli omologhi, durante la prima divisione meiotica ed è detto *crossing over*.

Durante lo stadio di sinapsi i due omologhi si trovano uno vicino all'altro. Quando sono vicini, essi raddoppiano i loro cromatidi, in modo da formare 4 filamenti, di cui due sono d'origine paterna e due d'origine materna; durante il processo di duplicazione, o immediatamente dopo, succede spesso che due di questi filamenti si scambino vicendevolmente delle porzioni, cioè per qualche ragione, un filamento materno e uno paterno si spezzano simultaneamente nello stesso punto. Le porzioni si risaldano reciprocamente attraverso un processo detto «ricombinazione»: in modo tale che risultano due nuovi cromatidi, ciascuno con una porzione di origine paterna e una di origine materna. Questo stadio particolare, riconoscibile al microscopio, è detto diplotene.

I punti in cui i cromatidi si scambiano, vengono detti punti di chiasma. Ciascuno dei 4 cromatidi così formati, ha la stessa probabilità degli altri di inserirsi in un gamete, di contribuire, cioè, al patrimonio genetico di un individuo. Ora, poiché ciascun cromatide può casualmente spezzarsi e riattaccarsi in un punto qualsiasi della sua lunghezza, la probabilità che uno dei figli erediti più geni che si trovano associati in uno dei genitori, risulta ulteriormente diminuita.

Questa stabilità nel numero e nella morfologia dei cromosomi non è però assoluta. Si conoscono vari modi, le diverse mutazioni cromosomiche, per mezzo delle quali i cromosomi di un individuo possono cambiare di numero e di forma. Abbiamo già accennato al caso di un cromosoma supplementare come quello della sindrome di Down, ma i cromosomi possono cambiare di forma per traslocazioni, inversioni, fusioni centriche o misdivisioni del centromero. Queste variazioni possono essere tali da differenziare il cariotipo di un individuo o di un gruppo di individui a tal punto da isolarli riproduttivamente dagli altri rappresentanti della loro specie. Non tutte queste variazioni tuttavia hanno possibilità di successo. Esse infatti per poter sopravvivere in una popolazione e per potersi affermare, devono presentare un qualche vantaggio selettivo e devono passare per il filtro della meiosi che si opera in ogni individuo, al momento della produzione dei gameti.

Fra i molti aspetti che interessano l'origine della nostra specie quello relativo alle trasformazioni del cariotipo ha stimolato molto interesse da quando Tijo e Levan nel 1956 riproposero con tecniche nuove il problema del numero e della morfologia dei cromosomi umani e da quando Lejeune nel 1959 individuò nella presenza di un cromosoma supplementare la causa della sindrome di Down.

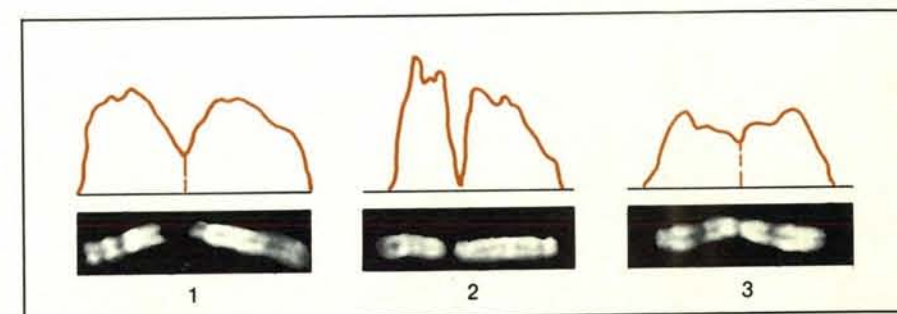
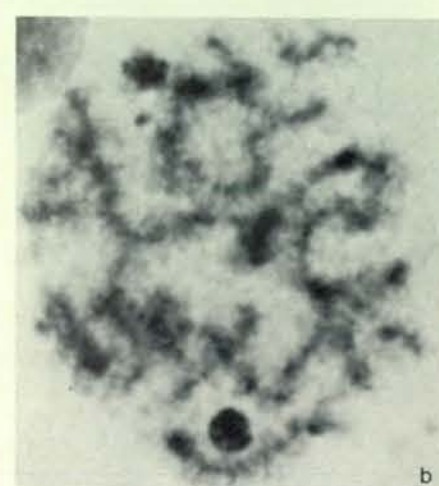
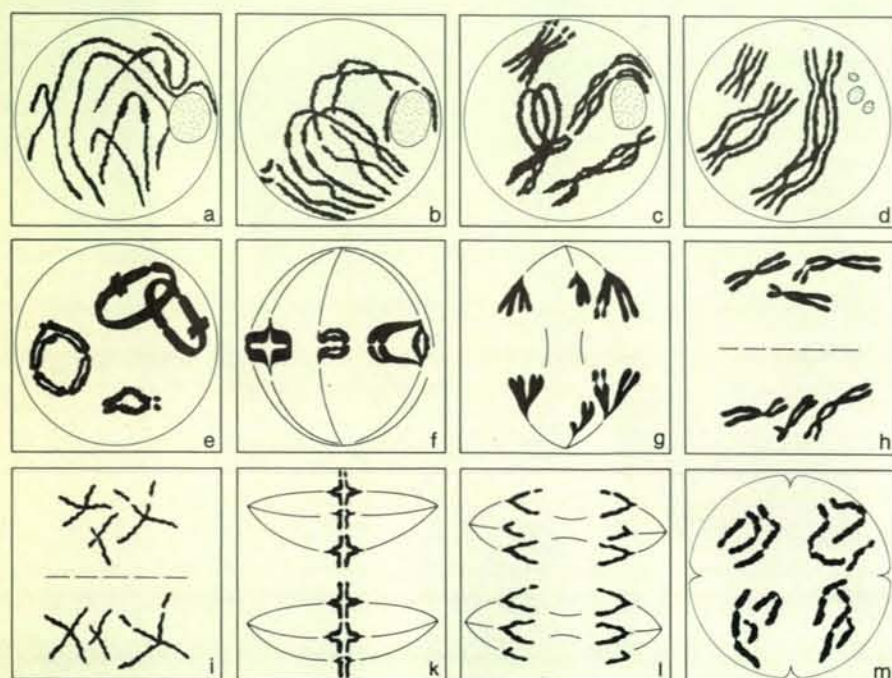
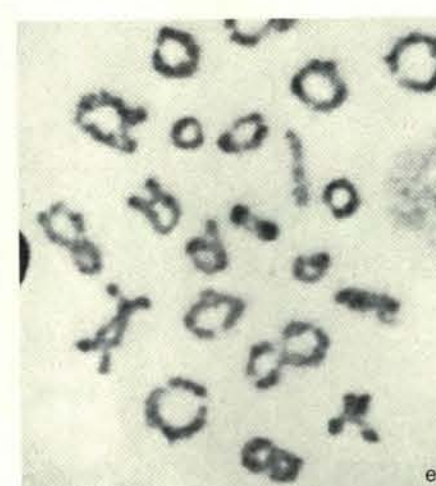
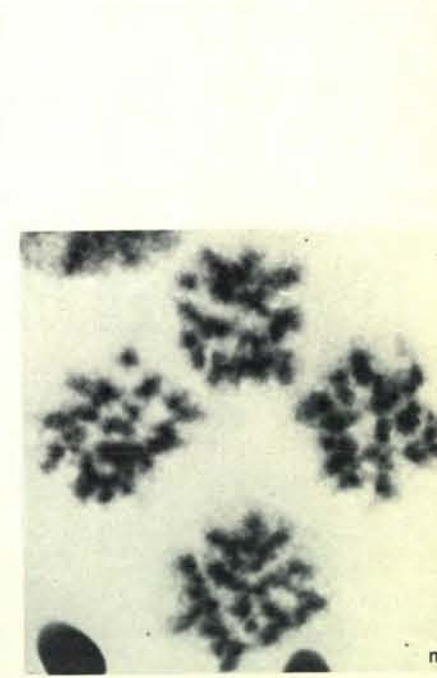
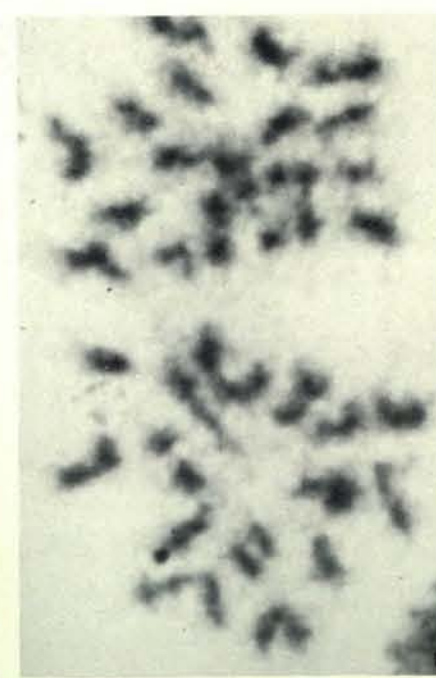
Queste strutture per essere i vettori della informazione ereditaria, hanno un indubbio interesse evolutivo oltre che costituire le fondamenta di un nuovo approccio alla patologia.

Tuttavia le aspettative di molti ricercatori sono state forse eccessive per una generalizzata tendenza a voler riconoscere nel cariotipo, il genotipo stesso e non il suo mero aspetto fenotipico. Questo ha indotto e induce costantemente alcuni medici biologi a trarre conclusioni fallaci e superficiali.

Per introdurre in modo adeguato il problema delle origini e della evoluzione del cariotipo umano occorre quindi fare alcune premesse sull'effettivo ruolo che le variazioni cromosomiche possono avere.

I cromosomi sono, come si diceva prima, delle mere strutture morfologiche e come tali a essi, per uno studio comparativo, devono essere applicati i criteri di omologia, analogia e convergenza. In biologia due strutture si definiscono omologhe quando hanno identica apparenza e medesima funzione, analoghe quando hanno la medesima funzione, ma diversa origine.

Nello studio comparativo dei cromosomi si devono quindi applicare i medesimi criteri che valgono per la comparazione delle strutture morfologiche, ma, a differenza di queste, lo studio comparato dei cromosomi presenta tre importanti vantaggi. Innanzitutto i cromosomi sono strutture relativamente semplici essendo formate fondamentalmente da una lunga catena di molecole di DNA; essi inoltre sono i vettori della informazione genetica e devono pertanto avere una morfologia costante in tutti gli individui della medesima specie in quanto alla meiosi queste

Questo cromosoma umano è stato fotografato al microscopio elettronico con un ingrandimento di 60 000 volte durante una delle fasi del processo di divisione del nucleo, la metafase.

Le tre fotografie rappresentano i cromosomi 1, 2 e 3 dello scimpanzè dopo trattamento con mostarda di quinacrina. In alto sono disegnate le rispettive curve di assorbimento. A ogni banda chiara del cromosoma corrisponde un massimo della curva, a ogni banda scura un minimo.

A sinistra sono state disegnate schematicamente le diverse fasi della divisione meiotica. In seguito a questa divisione si formano delle cellule con un patrimonio aploide di cromosomi equivalente alla metà di quello di una cellula somatica. I diversi stadi di queste divisioni sono: *a*) leptotene, *b*) zigotene, *c*) pachitene, *d*) diplotene, *e*) diacinesi, *f*) metafase I, *g*) anafase I, *h*) telofase I, *i*) interfase, *k*) metafase II, *l*) anafase II, *m*) telofase II. Nelle fotografie sono rappresentate alcune fasi caratteristiche come appaiono al microscopio ottico.

strutture si appaiano linearmente e pertanto la meiosi funziona da filtro per il controllo delle omologie. Va infine ricordato che si conoscono meccanismi relativamente semplici per mezzo dei quali i cromosomi possono cambiare di forma e di numero. Alcune volte questi mutamenti di struttura possono passare attraverso il filtro della meiosi ed eventualmente stabilizzarsi in una popolazione, come sequenze di informazioni organizzate in modo diverso.

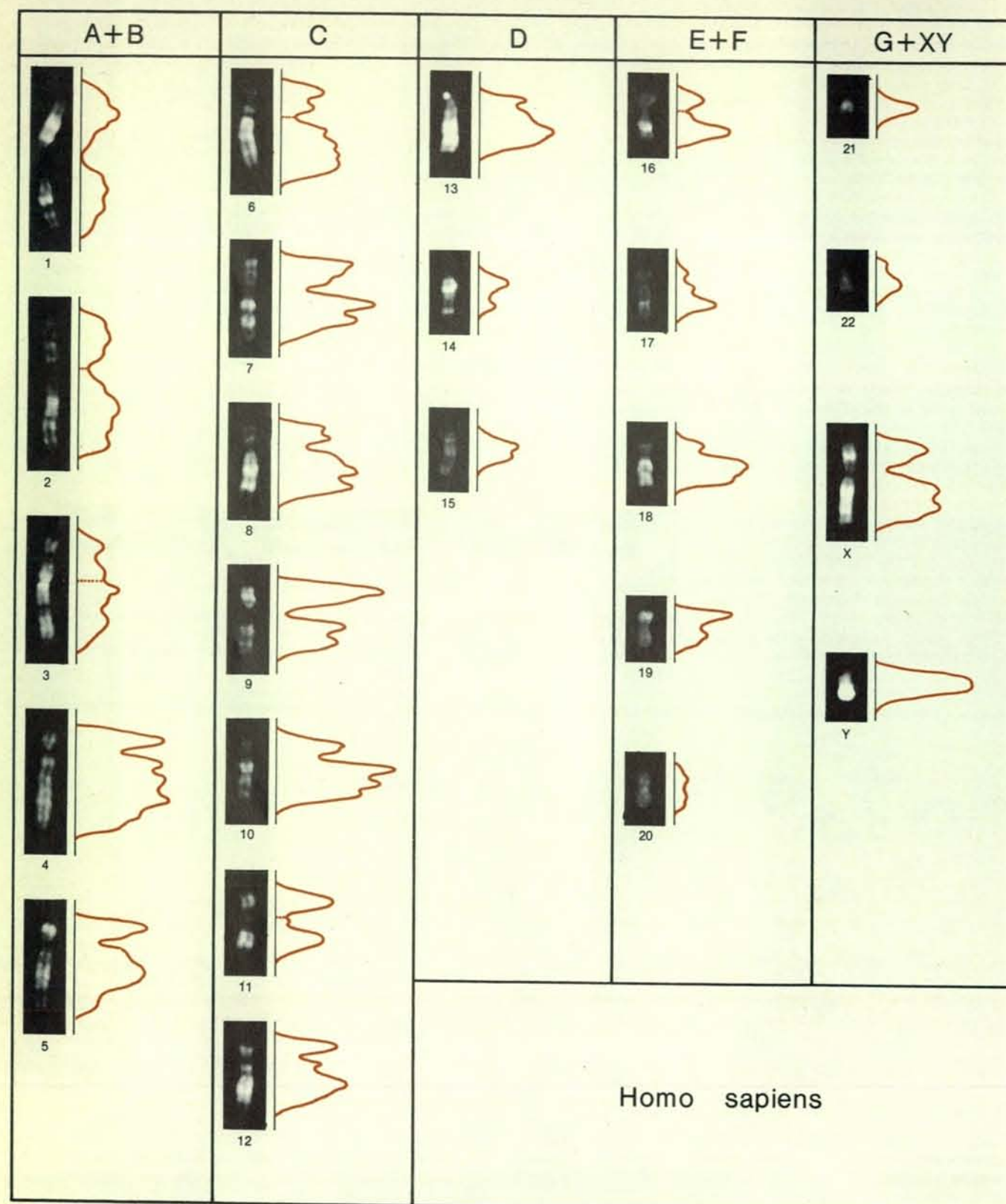
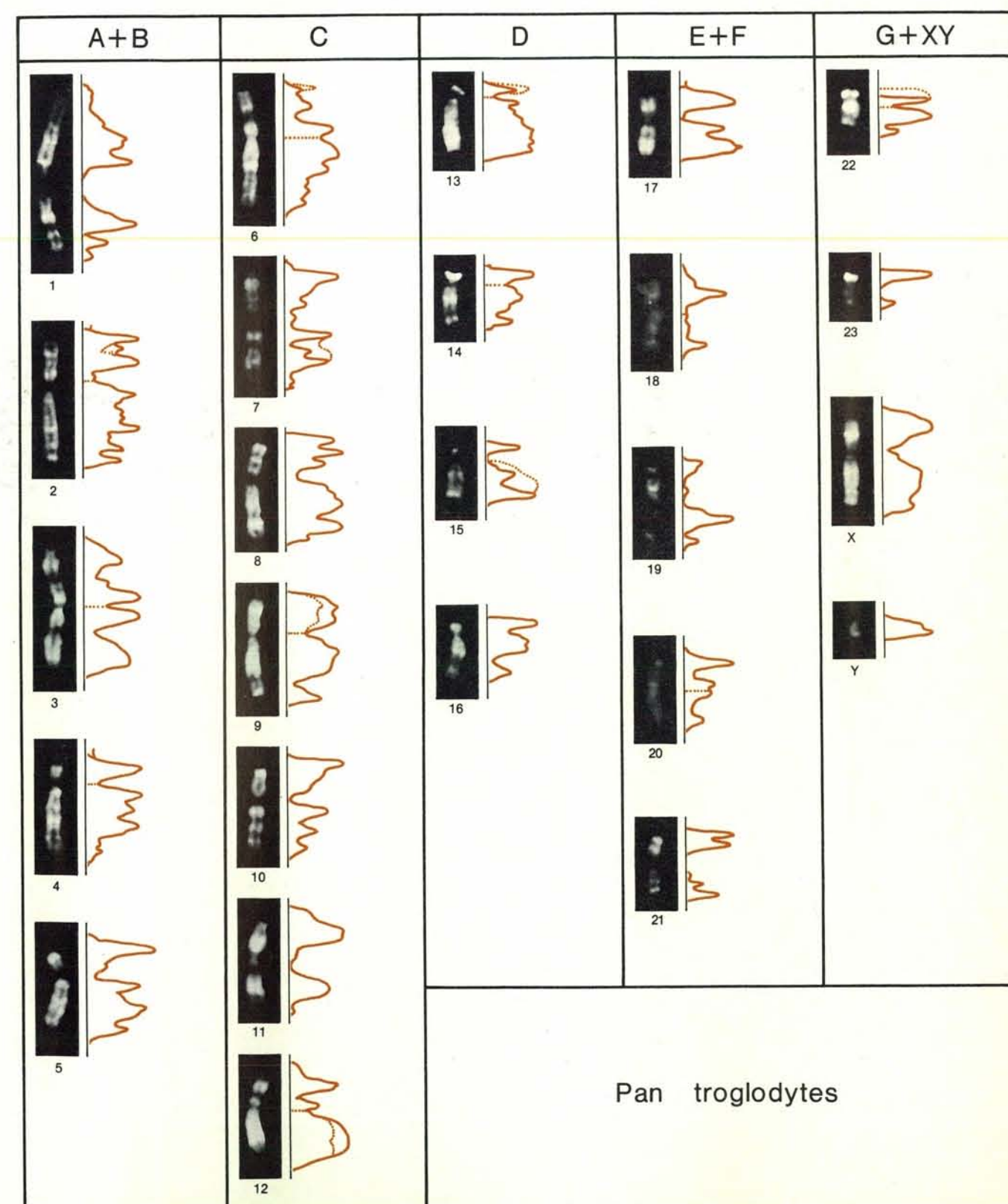
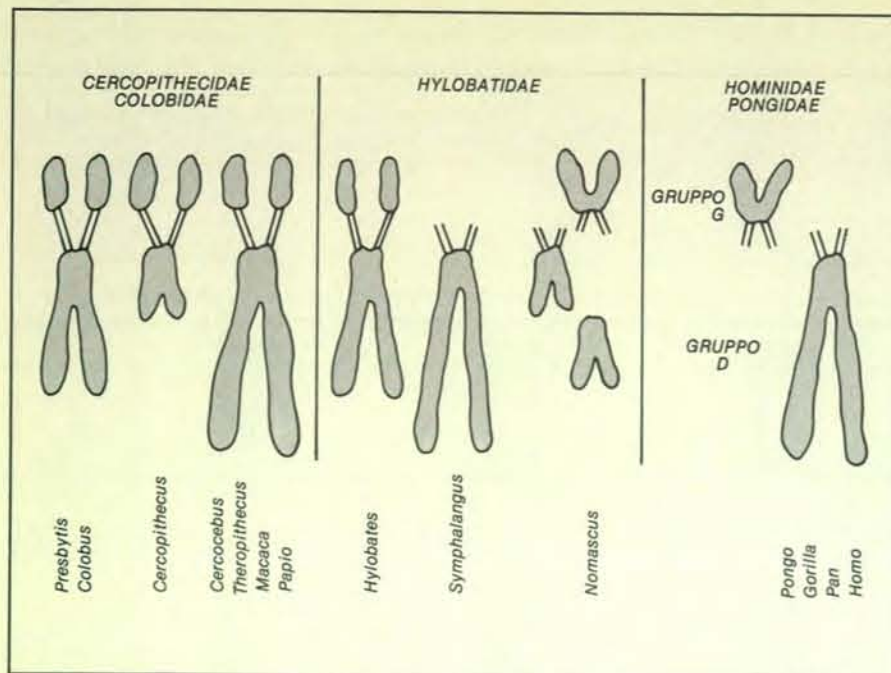
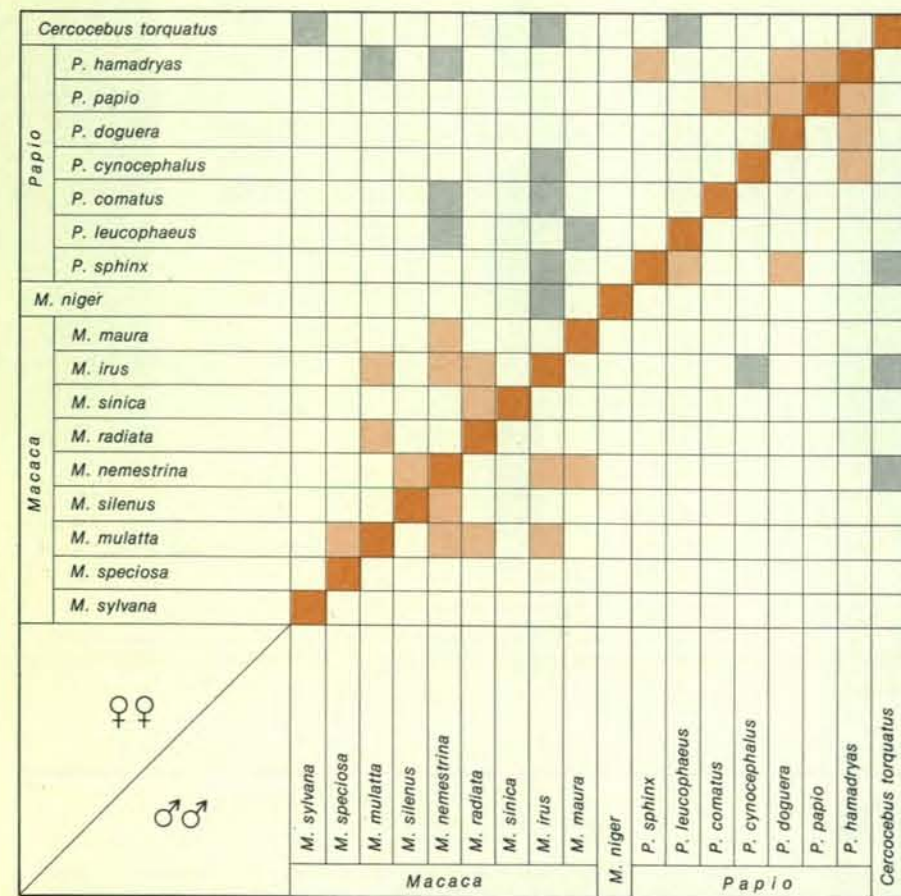
Per queste ragioni i cromosomi possono essere considerati ottimi caratteri per le ricerche sistematiche e filogenetiche.

La similitudine fra cromosomi di specie imparentate non implica tuttavia necessariamente la omologia genetica fra essi. Anche le più avanzate tecniche del bandeggiamento dei cromosomi, come vedremo, non forniscono prove inconfutabili di identità. Solo lo studio dei cromosomi meiotici di ibridi interspecifici può permettere di stabilire effettive omologie fra i complementi cromosomici di specie diverse.

Ma la scienza procede per gradi e noi dobbiamo molto spesso accontentarci di prove indirette, anche se plausibili, di omologia.

Qui di seguito cercheremo di sintetizzare i dati relativi alle possibili omologie fra i cromosomi umani e quelle delle specie a noi più strettamente imparentate e su queste basi potremo impostare il discorso della origine del cariotipo umano.

La nostra specie viene dagli zoologi classificata nell'infraordine delle scimmie catarrine e pertanto questo è il gruppo col quale più direttamente ci interessa comparare il cariotipo umano.

Se si confrontano per morfologia i ca-

Un accordo per la descrizione dei cromosomi umani è stato raggiunto a Denver nel 1960; nel 1963, 1966 e nel 1969 si sono tenuti altri convegni che hanno stabilito delle modalità universalmente accettate. Nelle figure di queste due pagine sono rappresentati i cromosomi umani disposti secondo la convenzione di Denver (*a sinistra*) e i cromosomi di scimpanzè disposti in un ordine analogo (*a destra*). I cromosomi sono stati colorati con quinacrina per metterne in evidenza le bande fluorescenti e per ciascuno è mostrata la relativa curva di assorbimento.

I cromosomi nucleolo organizzatori alla metafase presentano una evidente traccia di questa loro funzione in una ampia regione che non si colora con i coloranti tradizionali. Questi cromosomi che vengono detti «cromosomi marcati» rappresentano un ottimo indizio per stabilire possibili relazioni filogenetiche fra le specie. Nella figura è rappresentata una ipotetica ricostruzione delle tappe che hanno condotto alla evoluzione di questi cromosomi fra le scimmie catarrine e l'uomo.

Rappresentazione grafica dei possibili incroci che possono avvenire all'interno della famiglia dei cercopitecidi o scimmie cinomorfe. I quadretti in colore pieno rappresentano i normali incroci nell'ambito della stessa specie, quelli in colore chiaro gli incroci tra specie diverse mentre i quadretti simboleggiano gli incroci che è stato possibile osservare addirittura tra generi diversi.

riotipi delle diverse specie di catarrine si possono grosso modo distinguere quattro gruppi fondamentali: 1) il gruppo delle diverse specie del genere *Cercopithecus* caratterizzato dalla presenza di cromosomi metacentrici, submetacentrici e acrocentrici in proporzione varia; da una coppia di cromosomi acrocentrici marcata da una ampia regione acromatica e da un numero di cromosomi variabile fra 48 e 72; 2) il gruppo delle *Papinae* (*Macaca, Papio, Theropithecus*, e *Cercocebus*) caratterizzato dall'avere solo cromosomi metacentrici e submetacentrici, da una coppia di cromosomi submetacentrici marcata da una ampia regione acromatica e da un numero fisso di cromosomi pari a 42; 3) il gruppo delle *Colobinae* (*Colobus, Presbytis, Nasalis, Rynopithecus*) e delle *Hylobatinae* (*Hylobates, Nomascus* e *Symphalangus*) caratterizzati dall'avere cromosomi quasi tutti metacentrici in numero che va da 44 a 52 e da una coppia di cromosomi marcati da una ampia regione acromatica; 4) il gruppo delle scimmie antropomorfe (orango, gorilla, scimpanzé) e dell'uomo, caratterizzato dall'avere cromosomi metacentrici, submetacentrici e acrocentrici; dall'assenza di una coppia di cromosomi marcati e da un numero di 48 o 46 cromosomi.

È ovvio che la prima fase della ricerca di omologia fra il cariotipo umano e quello degli altri primati deve svilupparsi in questo ultimo gruppo.

La comparazione dei cariotipi delle scimmie antropomorfe con quello dell'uomo mette immediatamente in evidenza che il cariotipo dell'orango è il più differente, mentre quello dello scimpanzé è il più simile a quello dell'uomo. Anzi, a prima vista il cariotipo dello scimpanzé può essere facilmente scambiato con quello umano.

La differenza più appariscente è quella del numero. Lo scimpanzé, come le altre antropomorfe, ha 48 cromosomi, mentre l'uomo ne ha solo 46.

L'ipotesi più verosimile per spiegare queste differenze nel numero dei cromosomi è quella da noi avanzata fin dal 1962, di una fusione centrica avvenuta fra due cromosomi acrocentrici del tipo dei cromosomi del gruppo D dello scimpanzé che si deve essere verificata in un «preominide» ancestrale creando il cromosoma 2 del cariotipo umano.

Questa mutazione ha certamente prodotto una importante distinzione fra il *phylum* che ha dato origine alle scimmie antropomorfe e quello che ha dato origine all'uomo. Ma, a parte questa grossolana differenza numerica, molte altre differenze esistono fra il cariotipo di uno scimpanzé e quello dell'uomo. Alcune di queste differenze, probabilmente dovute a traslocazioni o a inversioni, hanno importanza forse anche maggiore di una semplice fusione centrica nel creare barriere riproduttive e quindi nuove popolazioni e specie.

Le possibili omologie di queste strutture fra le scimmie antropomorfe e l'uomo possono essere studiate con metodi diversi. Anche la semplice comparazione morfologica può dare un'idea approssimativa delle più grossolane differenze e di possibili omologie. Un certo numero di cromosomi a una ispezione morfodimensionale appaiono simili nelle quattro specie considerate. Attualmente tuttavia con le tecniche del bandeggiamento sembra possibile comparare anche solo porzioni di singoli cromosomi. Queste tecniche consistono nel trattare i cromosomi con agenti denaturanti o con quinacrina.

Mediante queste tecniche in genere risulta che i cromosomi 1, 3, 11, 12, 14 e X dello scimpanzé e dell'uomo sono caratterizzati da un medesimo tipo di bandeggiamento. Bandeggiamenti nettamente diversi rispetto agli omologhi umani invece li presentano i cromosomi dello scimpanzé 2, 4, 5, 8, 9, 16, 17, 18 e Y pur presentando caratteristiche morfodimensionali abbastanza simili a quelli umani. Essi quindi non sembrano avere una contropartе nel cariotipo umano. Gli altri cromosomi poi, come già abbiamo detto, sono nettamente diversi anche morfologicamente. Essi devono pertanto aver subito profondi rimaneggiamenti nell'una e nell'altra linea filetica.

È naturale che su questa base e con queste possibilità molti ricercatori si siano sentiti stimolati, in questi ultimi mesi, a inventare nuove tecniche per determinare differenze e similitudini fra il cariotipo umano e quello delle antropomorfe.

Ma quale è il significato di questi artefatti che noi produciamo sui cromosomi e che effettivo valore hanno per la comparazione di cromosomi di specie diverse? Prima di tentare di risolvere questi quesiti abbiamo voluto tentare di obiettivarne i risultati. Quando si usano le tecniche del bandeggiamento per appaiare due omologhi della medesima piastra o della medesima specie si può usare il principio di esclusione. Quando invece si comparano cromosomi di specie diverse, si deve essere certi che la identificazione delle bande sia ripetibile e obiettiva, cioè indipendente dalla acutezza visiva (che molto spesso è sostenuta dalla fantasia del ricercatore) e quindi sia operata da uno strumento.

Fra le diverse procedure messe a punto per produrre bande, il metodo che forse è più direttamente connesso con il DNA del cromosoma è quello della quinacrina, come ha dimostrato Caspersson nel 1972. Negli anni più recenti la topografia delle bande per fluorescenza da mostarda di quinacrina è stata studiata su specie differenti con risultati ottimi per il riconoscimento di cromosomi omologhi nell'ambito della medesima specie. Con il collega Giuseppe Ardito abbiamo quindi scelto questa tecnica per studiare comparativamente i cromosomi di uomo e di scimpanzé.

Nell'intento di ridurre le cause di variazioni dovute a differenze nella coltura e in altri procedimenti impiegati nell'approntamento dei cariotipi, abbiamo allestito colture miste di sangue di uomo e di scimpanzé. L'impiego di colture miste permette di trattare i cromosomi delle due specie nelle identiche condizioni eliminando tutte le differenze dovute ai passaggi nelle varie soluzioni e solventi. Le piastre delle due specie sul medesimo vetrino sono riconoscibili per il diverso numero dei cromosomi.

Al fine di obiettivare i risultati abbiamo poi effettuato la comparazione confrontando le curve ottenute per assorbimento fotoelettrico della fluorescenza dei cromosomi con la tecnica descritta da Caspersson nel 1973 per i cromosomi umani.

La comparazione diretta fra i cromosomi di uomo e di scimpanzé rivela alcune affinità nella successione delle bande più fluorescenti. Tuttavia, come dicevamo prima, la loro descrizione è opinabile sia per la dislocazione che per le dimensioni e l'intensità. L'uso dell'analisi fotodensitometrica avrebbe dovuto nei nostri programmi eliminare queste incertezze.

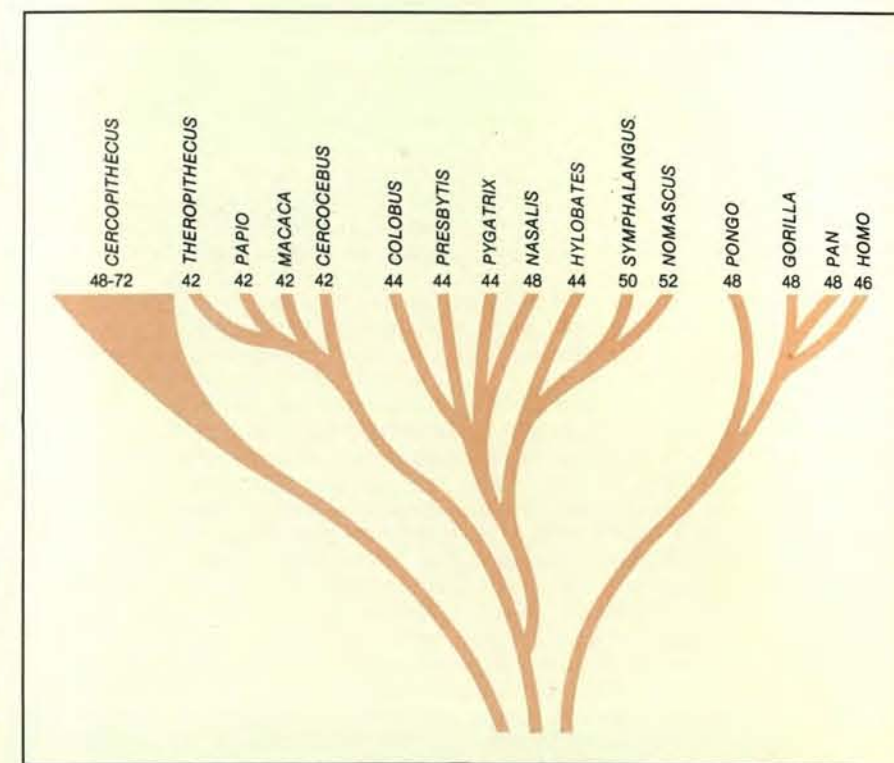
Questa tecnica, come abbiamo prima detto, si basa sull'impiego di un fotodensitometro che, con l'aiuto di un raggio luminoso, legge l'intensità della fluorescenza punto per punto lungo l'asse longitudinale del cromosoma. Questo strumento è connesso con una punta scrivente che registra l'intensità della fluorescenza sotto forma di una curva su una banda rotante. A ciascuna banda chiara corrisponde un massimo, a ciascuna banda scura un minimo.

La figura al centro a pagina 77 presenta le curve di assorbimento dei primi tre cromosomi di scimpanzé. Le curve dei due omologhi dei medesimi cromosomi si sovrappongono con notevole coincidenza. Abbiamo studiato le curve di assorbimento dei 23 cromosomi di un assetto aploide di uomo e dei 24 cromosomi di un assetto aploide di scimpanzé. Nonostante una generale somiglianza a prima vista, se si confrontano le curve dei possibili omologhi, è facile rilevare una costante eterogeneità che poco migliora i risultati osservati senza questa complessa procedura. Anche un tentativo di misurare le superfici delimitate da queste curve e di comparare le relative aree, ha dato risultati poco interessanti.

Questo significa che le tecniche del bandeggiamento non sono utilizzabili per individuare possibili analogie fra cromosomi di specie diversa? No, a nostro avviso il problema va risolto a monte.

Utilizzando queste metodiche è facile speculare e sviluppare facili ipotesi sulle varie tappe delle trasformazioni cromosomiche avvenute durante il processo di ominazione, ma è assolutamente pregiudiziale, prima di fare ciò risolvere i problemi di fondo sulle cause per le quali i trattamenti utilizzati producono queste alterazioni strutturali lungo i cromosomi poiché diverso è il loro significato se coinvolgono il DNA o le proteine che rivestono il cromonema.

Ma bande o non bande, possibili omologie fra i cromosomi dell'uomo e degli altri primati possono essere dedotte anche considerando analogie fisiologiche e omologie geniche.

Per questo tipo di ricerche speciale interesse hanno i cosiddetti cromosomi marcati che, come abbiamo visto, sono

L'analisi cromosomica permette una ricostruzione filogenetica dei primati che in larga misura coincide con quella tradizionale, tuttavia alcune informazioni la migliorano e la superano. Nella figura è rappresentato un albero filogenetico su base cariologica dei primati del Vecchio Mondo (scimmie catarrine) e delle relazioni che fra essi intercorrono. I tre rami non sono per ora accomunabili, ma i cromosomi marcati possono rappresentare un elemento di connessione.

DATI QUANTITATIVI		GENERE	DATI QUALITATIVI				
LUNGHEZZA TOTALE DEI CROMOSOMI (MICROMETRI)	NUMERO CROMOSOMICO DIPLOIDE		TIPO DI CROMOSOMA			CROMOSOMA MARCATO	CROMOSOMA Y
			M	S	A		
92 ± 12	42	*Macaca*	6	13	—	*A*	*b*
90 ± 10	42	*Cynopithecus*	6	13	—	*A*	*b*
88 ± 10	42	*Papio*	6	13	—	*A*	*b*
89 ± 10	42	*Theropithecus*	6	13	—	*A*	*b*
85 ± 10	42	*Cercocebus*	6	13	—	*A*	*b*
94-125 ± 10	54-72	*Cercopithecus*	6-9	12-17	6-10	*C*	*a,b,c,*
94 ± 10	54	*Erythrocebus*	6	12	7	*C*	*b*
—	—	*Pygathrix*		—		—	—
—	—	*Rhinopithecus*		—		—	—
—	—	*Simias*		—		—	—
—	48	*Nasalis*	8	15	—	*A*	—
83 ± 10	44	*Presbytis*	7	12	1	*B*	*a*
93 ± 10	44	*Colobus*	7	13	—	*B*	—
85 ± 10	44	*Hylobates*	11	9	—	*B*	*c*
—	50	*Symphalangus*	12	11	1	—(?)	*c*
83 ± 10	48	*Pongo*	—	12	11	—	*a*
94 ± 10	48	*Pan*	5	10	8	—	*a*
98 ± 10	48	*Gorilla*	5	10	8	—	*a*
93 ± 10	46	*Homo*	4	13	5	—	*a*

Dati cariologici quantitativi e qualitativi per una revisione tassonomica a livello sopragenerico dei primati del vecchio mondo (scimmie catarrine). I dati quantitativi delle colonne di sinistra concernenti il contenuto in DNA per nucleo e la lunghezza totale dei cromosomi sono ovviamente indicativi poiché soggetti a errori strumentali. Per quanto concerne i dati qualitativi con *M* si sogliono indicare i cromosomi metacentrici, con *S* i submetacentrici e con *A* i cromosomi acrocentrici. Con il termine cromosomi marcati si indica una coppia di cromosomi presente in molte specie di questo gruppo e caratterizzata da un'ampia regione acromatica su uno dei bracci; le lettere *a*, *b* e *c* indicano rispettivamente tre differenti tipi di questi cromosomi: *a*) con braccia opposte a quelle con regioni acromatiche lunghe, *b*) con braccia opposte alla regione acromatica di media lunghezza, *c*) con braccia opposte a quelle aventi regione acromatica corta o quasi inesistente.

caratteristici di quasi tutti i primati del vecchio mondo.

Questi cromosomi sono innanzitutto interessanti per il loro sistema di «reduplicazione». Huang, Habbitt e Ambrus nel 1969 hanno dimostrato che in *Macaca mulatta* questi cromosomi si reduplicano asincronicamente rispetto agli altri cromosomi. Questa osservazione indica che, nel paio di cromosomi marcati, la sintesi del DNA si completa prima che nel restante corredo cromosomico. Il fatto che questo cromosoma presenti una ampia costrizione secondaria indicherebbe che esso è il sito della informazione del nucleolo il quale è, come noto, il centro della sintesi del RNA ribosomale. Il precoce completamento della sintesi del DNA in un cromosoma, associato alla formazione del nucleolo, sembra essere vantaggioso per l'economia della cellula. Completata la sintesi del DNA, questa regione può iniziare la riorganizzazione del nucleolo nella telofase e può quindi esplicare la sua funzione prima che il DNA si despiralizzi nella interfase successiva.

Ma un altro aspetto rende interessanti questi cromosomi. Nell'uomo le regioni nucleo-organizzatrici sono situate nelle costrizioni secondarie che uniscono i satelliti alle braccia corte dei cromosomi acrocentrici (13-15, 21-22). Ovviamente una comparazione diretta di questi cromosomi con i cromosomi nucleolo organizzatori delle *Papinae* è puramente speculativa. Tuttavia la somiglianza, sia nella forma che nella dimensione, che risulterebbe dalla ipotetica associazione di questi due tipi di cromosomi nell'uomo e nelle scimmie antropomorfe con il cromosoma marcato, è notevole. A parte l'interesse genetico, la prospettiva di questa omologia offrirebbe un nuovo approccio nell'affrontare il problema del cariotipo ancestrale da cui avrebbero avuto origine i vari gruppi delle scimmie del vecchio mondo. Infatti nella interpretazione filogenetica da noi prospettata nella figura a pagina 81 si riconoscono tre linee evolutive, ma nessuna possibile connessione è prevista tra di loro.

Informazioni determinanti sulla possibile omologia genetica di un particolare cromosoma nell'uomo e nello scimpanzé sono state ottenute da McClure e dai suoi collaboratori dello Yerkes Primate Center di Atlanta. In questo centro di primatologia nel 1970 nacque uno scimpanzé che presentava un piccolo cromosoma acrocentrico soprannumerario. Questo scimpanzé neonato presentava molte anomalie fenotipiche simili a quelle di individui umani affetti da mongolismo o sindrome di Down i quali appunto hanno una trisomia del cromosoma 21. L'omologia, da un punto di vista morfologico, tra il cromosoma 21 dell'uomo e la tripletta presente in questo scimpanzé dello Yerkes è sorprendente. Ne consegue che l'omologia genetica perlomeno parziale di questo cromosoma fra l'uomo e lo scimpanzé può essere ragionevolmente sostenuta.

Non c'è dubbio che in un prossimo futuro si possano ottenere, con relativa facilità, informazioni attendibili sull'omologia genetica dei cromosomi X delle diverse specie di primati. Il cromosoma X sembra essere abbastanza stabile per il suo contenuto genetico nelle varie specie di mammiferi e già per la specie umana si conoscono numerosi marcatori e molti sono stati descritti per diverse specie di primati non umani.

La omologia di informazioni genetiche fra i cromosomi Y può poi essere dedotta dal prodotto diretto del loro genotipo: gli spermatozoi. In una ricerca al microscopio a scansione, recentemente pubblicata sul «Journal of Human Evolution», Martin e collaboratori hanno dimostrato inequivocabilmente la stretta similitudine fra la morfologia degli spermatozoi di uno scimpanzé con quelli dell'uomo, mentre quelli di altri primati appaiono considerevolmente diversi.

Lo sviluppo recente di ibridi cellulari per decifrare la sequenza dei geni sui cromosomi umani sta offrendo nuove e forse più concrete prospettive per individuare l'esistenza di omologie genetiche.

In che modo i differenti eventi mutativi che hanno così profondamente differenziato il cariotipo dell'uomo rispetto a quello delle altre antropomorfe si sono affermati? È eventualmente possibile stabilire il tempo e la sequenza della loro comparsa?

Per comprendere come queste mutazioni possono essersi affermate in un preominide ancestrale è necessario innanzitutto considerare la dimensione demografica dei nuclei comunitari che costituivano le popolazioni di allora. Essi non dovevano superare la ventina di individui. Erano «clan» familiari molto spesso dominati da un solo maschio che godeva del dominio assoluto sulle femmine del gruppo. Neppure doveva essere infrequente che alcuni maschi si accoppiassero con le proprie figlie. La *leadership* in un gruppo di scimpanzé infatti sembra durare anche più di dieci anni e la maturità sessuale della femmina è raggiunta a 8 anni.

Un maschio portatore di una fusione centrica, cioè con un cariotipo di 47 cromosomi, per esempio, aveva quindi due possibilità di affermare questo nuovo cariotipo completando la riduzione del numero dei cromosomi in 46: quella di accoppiarsi con le varie femmine del gruppo, producendo così prole per metà con 47 cromosomi e dando quindi alla generazione successiva la possibilità di completare questa riduzione; oppure quella di accoppiarsi con le sue figlie mediante una sorta di reincrocio. In questo caso il processo di riduzione del cariotipo sarebbe stato ancora più rapido e potrebbe essere avvenuto nel volgere di una decade. In ogni caso non avrebbe coinvolto più di due generazioni, al massimo quindi trenta anni, considerando il più breve tempo di acquisizione della maturità sessuale in questi nostri antenati.

La possibilità di affermazione di una tale mutazione, è ovviamente condizionata, oltre che dall'esistenza di piccole comunità riproduttive, anche dal potere intrinseco di affermazione che questa mutazione deve aver posseduto.

Ovviamente poco si può dire a questo livello, tuttavia il blocco di più geni in un medesimo cromosoma può aver rappresentato un qualche vantaggio selettivo in un certo momento della evoluzione degli ominidi.

Ma con queste premesse potremo porci altre domande. È il cariotipo attuale dell'uomo in equilibrio stabile? In altre parole rappresenta esso il massimo vantaggio per la nostra specie o può ulteriormente variare? Se, per esempio, non avessimo cromosomi acrocentrici del tipo 21 e se questi fossero collegati mediante fusione centrica con l'altra coppia di cromosomi acrocentrici del gruppo G o con quelli del gruppo D ottenendo un cariotipo con 44 cromosomi non si avrebbero più sindromi di Down con grande vantaggio per la specie. Appare quindi evidente che il cariotipo umano può subire ulteriori miglioramenti.

L'esame esteso al livello di intere popolazioni umane ha permesso di individuare casi di fusione centrica fra cromosomi acrocentrici con la formazione di cariotipi di 45 cromosomi in individui apparentemente normali.

Se due di questi individui si accoppiassero, si otterrebbero figli con 44 cromosomi probabilmente perfettamente normali. Ma in una popolazione attuale questo è difficile che avvenga: individui con cariotipo di 45 cromosomi per fusione centrica sono piuttosto rari e una mutazione del genere viene dispersa facilmente. Ma la possibilità sussiste.

Molto probabilmente tuttavia il nostro considerare i cromosomi come unità indipendenti è errato. A nostro avviso esistono altre forze che coordinano e in qualche modo mantengono stabile il cariotipo di una specie.

È probabile infatti che i cromosomi di ciascun gamete siano concatenati tra loro a formare un grande anello. Vi sarebbero pertanto due di questi anelli per ciascuna cellula somatica: uno di origine paterna e uno di origine materna. Ma le conseguenze citotassonomiche di questo discorso ci porterebbero troppo lontano.

Un altro modo per stabilire le reazioni filogenetiche tra le varie specie di primati è la comparazione diretta del contenuto in DNA dei nuclei delle cellule delle diverse specie.

Dati quantitativi sono stati di recente ottenuti e accuratamente elaborati da Maria Gabriella Manfredi Romanini e dai suoi collaboratori dell'Istituto di antropologia dell'Università di Pavia. Ma i dati quantitativi hanno un interesse limitato anche perché in molte specie esiste una certa quantità di DNA altamente ripetitivo che tende a far svalutare questo tipo di informazione.

Più interessante quindi è la comparazione diretta delle sequenze delle basi puriniche e pirimidiniche (adenina, citosina, guanina, timina) nel DNA nucleare delle diverse specie in quanto direttamente coinvolte col cosiddetto codice genetico.

Questo tipo di comparazione è ora possibile mediante tecniche relativamente semplici. Il principio si basa sulla dissociazione della doppia catena del DNA mediante la rottura dei ponti idrogeno fra le due catene complementari e la riassociazione di queste singole catene provenienti da DNA di specie diverse. Questo DNA ibrido ci informerà sulla qualità di sequenze omologhe ed eterologhe presenti in una specie rispetto all'altra.

Praticamente si procede alla denaturazione del DNA (separazione della doppia elica nei due filamenti) mediante ebollizione in soluzione salina facendo riassociare porzioni di catene di 400-500 nucleotidi con filamenti di DNA di altre specie marcati con timidina radioattiva.

Durante questa fase di riassociazione i filamenti tendono a riassociarsi in quantità variabili a seconda della quantità di basi complementari. Vi saranno pertanto dei filamenti perfettamente accoppiati, dei filamenti più o meno accoppiati e dei filamenti liberi. La separazione fra i filamenti accoppiati e quelli rimasti liberi la si ottiene facendo passare la miscela attraverso una colonna di idrossiapatite. Questo cristallo ha la proprietà di adsorbire le molecole di DNA doppie o comunque riassociate, ma non i singoli filamenti che possono essere asportati con il semplice lavaggio. Eluendo poi a temperature gradualmente incrementate, sui cristalli di idrossiapatite si dissocieranno le coppie di filamenti a cominciare da quelli a minor numero di basi complementari. Le frazioni eluite vengono determinate di volta in volta quantitativamente utilizzando la loro radioattività, mediante un contatore a scintillazione. La stabilità termica del DNA riassociato infine può essere confrontata con quella del DNA naturale che possiede accoppiamenti perfetti fra le basi complementari.

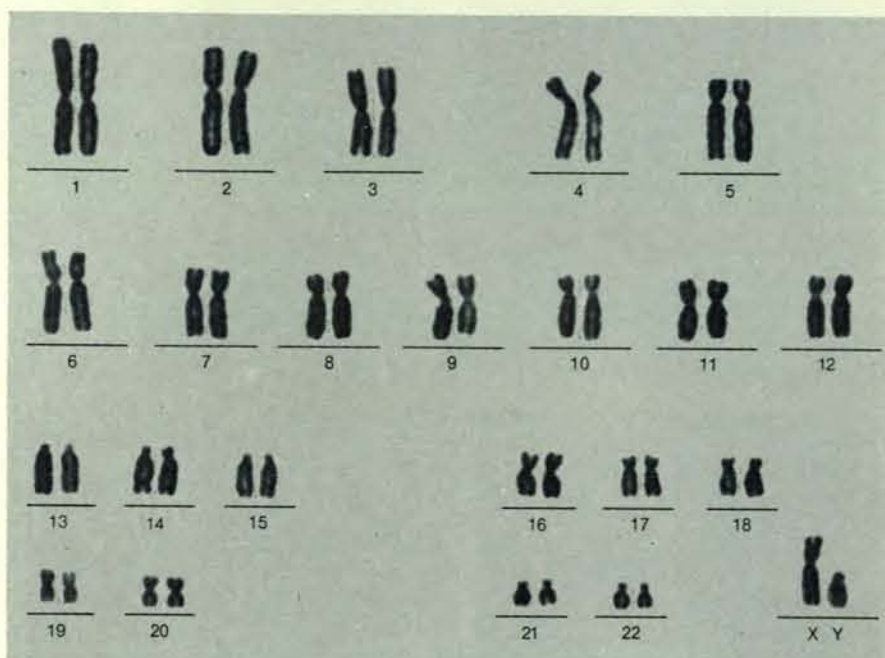
Questo procedimento è stato applicato da Kohne in un recente articolo pubblicato su «Journal of Human Evolution» nei confronti di diverse specie di primati in relazione al DNA umano con risultati molto interessanti. Secondo questi dati la percentuale di nucleotidi diversi rispetto all'uomo sarebbe del 2,4 per cento con lo scimpanzé, del 5,3 per cento con il gibbone, del 9,5 per cento con il cercopiteco, del 15,8 per cento con il cebo e di circa il 42 per cento con il galagone. Questi dati dimostrano una incrementata diversificazione con la distanza biologica fra le specie e sono in generale accordo con la sistematica tradizionale.

Questi dati poi possono essere utilizzati per la costruzione di un albero filogenetico basato direttamente sul numero delle mutazioni intercorse durante la differenziazione delle specie, anche se per questo si devono tenere nelle dovute considerazioni due limitazioni. In un primo luogo una grande quantità di DNA presente nel nucleo delle diverse specie è, come si è detto, di tipo ripetitivo, cioè presenta una grande quantità di basi ripetute ed è forse costante fra specie anche molto differenti. In secondo luogo, mentre per il calcolo delle mutazioni fra specie affini le sostituzioni dei nucleotidi sono quelle avvenute durante la evoluzione delle specie, per quelle più lontane i valori sono inferiori al reale a causa delle sostituzioni multiple avvenute nello stesso punto e che pertanto non è possibile scoprire.

Una terza limitazione va cercata nella influenza che ha la differenza nella lunghezza del tempo di generazione fra le specie che si intendono comparare; il tempo di generazione di un uomo è di 20 anni, quello dello scimpanzé di 12 mentre quello di un galagone di soli 2 anni.

Nonostante queste limitazioni è stato tentato il calcolo dei tempi di divergenza fra l'uomo e diverse altre specie di primati con risultati attendibili. Essi sono: uomo-scimpanzé 15 milioni di anni, uomo-gibbone 30, uomo-cercopiteco 45, uomo-cebo 65, uomo-galagone 80.

Ma tutte le variazioni di DNA sono dovute a questo tipo di meccanismi casuali o altri fenomeni interagiscono nel produrre variazioni nel DNA di una determinata specie? A più riprese sono sorti dubbi e incertezze. La recentissima scoperta di frazioni di DNA identiche fra l'uomo, il babbuino e il gatto, fatte da George Todaro del National Cancer Institute di New York estenderebbe agli animali superiori meccanismi di inserzione genetica noti per i batteri. Una prova, ancora, dell'esistenza di una matrice unica che governa l'organizzazione degli esseri viventi dal batterio all'uomo.

I cromosomi, dopo essere stati colorati con coloranti opportuni, sono fotografabili al microscopio ottico. Le fotografie vengono poi notevolmente ingrandite per il ritaglio e l'appaiamento dei cromosomi omologhi. Come mostra la fotografia in alto, i cromosomi sono formati da due cromatidi ravvicinati uniti in un punto caratteristico detto centromero.

La figura nella pagina a fronte mostra come le 24 coppie di omologhi dei cromosomi umani, a eccezione dei cromosomi sessuali, vengono disposte in ordine di dimensione e ordinate in base alla posizione del centromero che permette di distinguere i cromosomi in acrocentrici, metacentrici, e submetacentrici.

Le tecniche di bandeggiamento introdotte nel 1971 consistono nel trattare i cromosomi fissati con agenti denaturanti di varia natura (quinacrina, NaOH, calore, tripsina ecc.). Questo trattamento provoca sul cromosoma la comparsa di bande chiare e scure le quali permettono il riconoscimento inequivocabile degli omologhi. La piastra qui fotografata rappresenta dei cromosomi umani metafasici.

Ragni sociali

Quasi tutti i ragni conducono vita solitaria; alcune specie, tuttavia, sono gregarie e altre costruiscono addirittura grandi ragnatele comuni. In alcune specie messicane si osservano entrambi i gradi di socialità

di J. Wesley Burgess

Tra gli insetti, in particolare tra le api, le formiche e le termiti, il comportamento sociale è molto comune, mentre tra i ragni è raro. Tutti i ragni sono carnivori predatori; in molte specie il maschio non può addirittura avvicinarsi alla femmina senza correre il rischio d'essere attaccato e ucciso. Sembra perciò un paradosso che possano esistere specie di ragni sociali.

Il numero di ragni sociali è piccolo: solo 12 generi distribuiti in 9 famiglie presentano casi di socialità. Tuttavia i 12 generi presentano un'area di distribuzione assai estesa, con rappresentanti sia nel nuovo sia nel vecchio mondo. Due delle specie del nuovo mondo si trovano in Messico. Recentemente ho visitato zone nei pressi di Guadalajara dove entrambe le specie sono presenti, ho osservato i ragni sociali nel loro habitat naturale e ho catturato un certo numero di esemplari per allevarli e osservarli nel mio laboratorio della North Carolina State University.

Le due specie messicane hanno una vita chiaramente diversa. *Mallos* (un tempo chiamato *Coenothele*) *gregalis* è un ragnetto il cui corpo raramente supera la lunghezza di 5 millimetri: costruisce una grande ragnatela comune, circondando i rami d'un albero con uno strato continuo di seta. Le sue colonie possono essere considerate, dal punto di vista sociale, le più complesse di tutta l'America settentrionale. *Oecobius civitas* è un ragno ancora più piccolo: pochi sono gli individui che presentano un corpo più lungo di 2 millimetri e mezzo. Questa specie vive a gruppi, e fila la sua tela di seta a scopo di rifugio e di sistema d'allarme in un microhabitat scuro e ristretto: fessura di una roccia.

Le società dei ragni sono diverse da quelle formate dagli insetti sociali, sia per il tipo, sia per il grado di socialità. Una delle ragioni di questa diversità è che la tela di un ragno estende l'ambito della percezione sensoriale in un modo che non ha analogia nel mondo degli insetti. Un'altra ragione sta nel fatto che un ragno possiede parti boccali conformate in modo che può nutrirsi solo di altri organismi animali. Qualsiasi animale di dimensioni appropriate che un ragno può incontrare, compreso un ragno di un'altra specie o persino della medesima specie, è una preda potenziale. Comunque, ci sembra utile descrivere la socialità dei ragni per tratteggiare l'evoluzione probabile dei diversi gradi di socialità tra gli insetti.

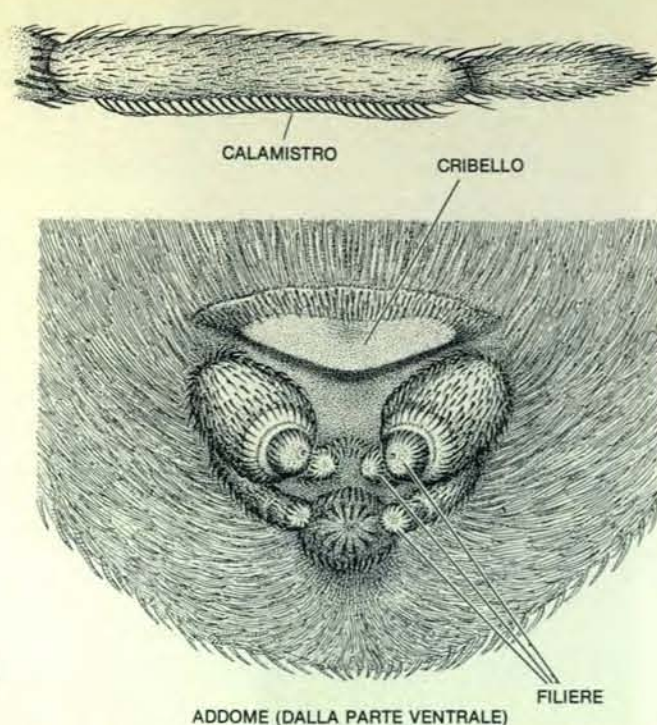
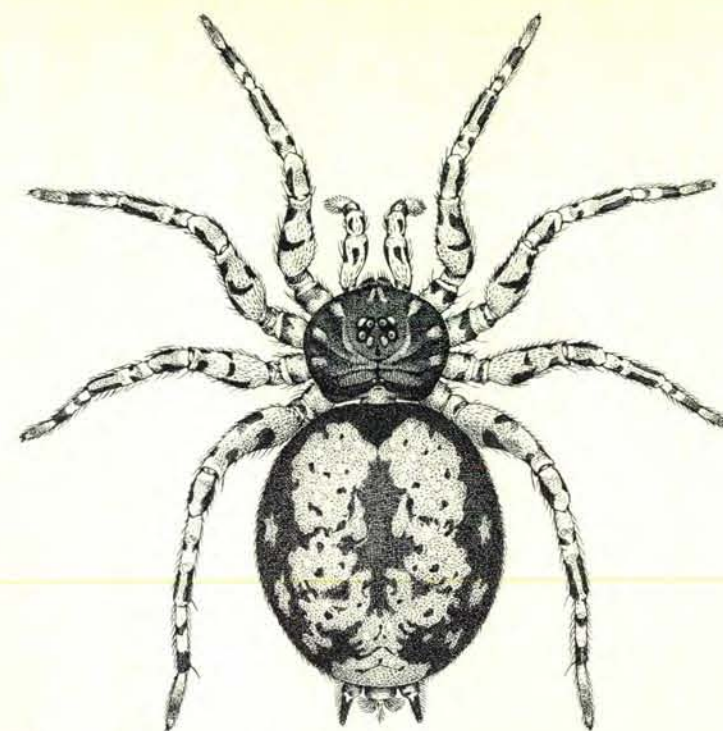
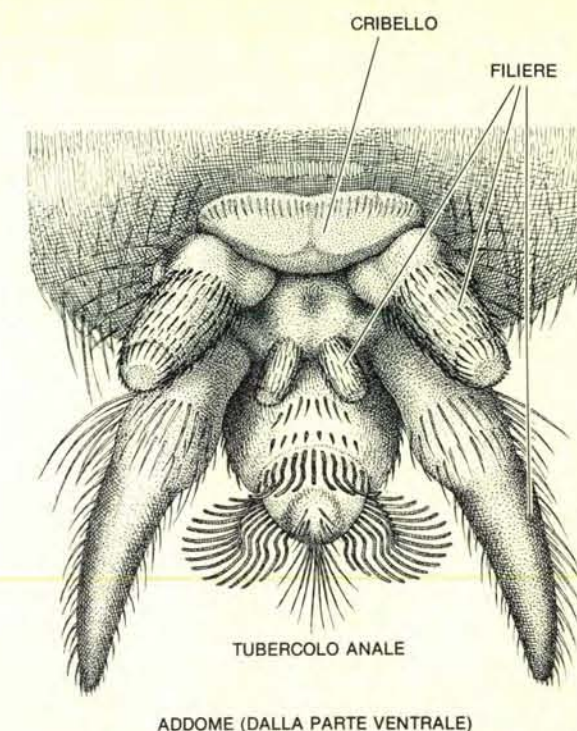
Come già Edward O. Wilson, della Harvard University ha fatto notare, gli insetti «eusociali», ossia gli insetti sociali di grado più evoluto, hanno tre caratteristiche in comune: cura cooperativa dei giovani, divisione del lavoro, per cui gli individui più o meno sterili si occupano delle necessità degli individui fecondi, e un ciclo vitale abbastanza lungo affinché la prole possa a un certo punto condividere le attività della generazione parentale. Sembra che le vie evolutive che possono aver condotto gli insetti da un tipo di comportamento non sociale a quello eusociale siano rintracciabili nel comportamento primitivo rispetto a quello eusociale, che si riscontra in parecchi insetti affini alle specie eusociali. Charles D. Michener dell'Università del Kansas ha delineato due possibili vie evolutive.

La prima via viene definita da Michener «parasociale»: in essa si possono indicare tre livelli di comportamento sempre più complesso sulla via dell'eusocialità. Il livello inferiore, ossia il comportamento comunitario, è caratterizzato da un'aggregazione di femmine, tutte appartenenti alla medesima generazione: appena queste si sono aggregate, costruiscono un nido comune per i loro piccoli. Il livello successivo, detto comportamento quasi-sociale, è caratterizzato dalla cura cooperativa dei piccoli. Il terzo livello, o comportamento semisociale, è caratterizzato dall'apparizione di diverse caste, che possiedono funzioni specifiche differenti. Dopo di ciò l'eusocialità viene raggiunta quando il ciclo vitale si allunga, cosicché i genitori e la prole che ha raggiunto la maturità coesistono nella medesima colonia.

La seconda via evolutiva possibile viene definita da Michener «subsociale». Su questo tragitto evolutivo esiste solo un livello di comportamento che precede l'eusocialità: è caratterizzato da una costruzione di nidi solitaria, piuttosto che comunitaria. Tuttavia la femmina solitaria rimane nel nido e prende cura dei piccoli. L'eusocialità viene raggiunta quando la costruttrice del nido vive abbastanza a lungo per ricevere l'assistenza della sua prima generazione figlia nella cura delle successive generazioni.

Considerato sotto questo aspetto, non esiste alcun ragno sociale che si possa

Nella fotografia della pagina a fronte si può osservare la cattura d'una mosca (*A*) eseguita in cooperazione da parecchi ragni. Il disegno schematico qui sopra serve per identificare le prede e i predatori. I ragni indicati con *M* sono maturi, quelli indicati con *I* sono immaturi. Solo due delle mosche che appaiono nella tela (*A e B in basso a sinistra*) rappresentano prede fresche. I ragni appartengono alla specie sociale *Mallos gregalis*. Il gruppo di ragni maturi sta mangiando o preparandosi a mangiare. Un ragno immaturo è stato attirato sulla scena e un altro si sta avvicinando. La fotografia è stata eseguita in laboratorio dall'autore; i ragni sono stati raccolti in Messico.

Il ragno sociale *Mallos gregalis* ha un corpo della lunghezza di 5 millimetri. La sua ragnatela complicata, che funziona come trappola per le mosche, comprende molte fasce appiccicose di seta che invischiano gli intrusi. La seta appiccicosa viene filata attraverso centinaia di fori microscopici in una piastra addominale (*a destra*), detta cribello. Il ragno pettina la seta con un organo chiamato calamistro.

Il ragno gregario *Oecobius civitas* ha un corpicciolo della lunghezza media di 2,5 millimetri. Come *Mallos gregalis*, possiede un cribello, ma utilizza la seta appiccicosa attivamente, avvolgendo la preda anziché aspettare che questa resti invischiata. Il ragno produce la seta con le filiere e la pettina mediante il tubercolo anale (*a destra*) utilizzandola per avvolgere le sue prede (*si veda la figura a pagina 93, al centro*).

definire eusociale. Dobbiamo delimitare la base comune della socialità dei ragni sotto aspetti molto più ristretti: esistenza di vari gradi di comportamento comunitario e di interazioni tra i membri delle aggregazioni comunitarie.

A questo punto si dovrebbe notare che con poche eccezioni persino i ragni che hanno abitudini solitarie attraversano uno stadio semicomunitario già nelle prime fasi della loro esistenza. A differenza degli insetti, infatti, i ragni non passano attraverso uno stadio larvale: ciascuno emerge dall'uovo come un adulto in miniatura, pur conservando un sacco vitellino che lo rifornisce di sostanze nutritive per parecchi giorni; cresce di dimensioni e sviluppa le sue caratteristiche sessuali attraverso una serie di mute successive, la prima delle quali avviene al riparo del sacco ovigero parentale: quando abbandona il sacco ovigero è già in grado di tessere il filo di seta e di metter fuori combattimento le prede.

Ci si aspetterebbe quindi che i ragnetti delle specie solitarie si sparpaglino non appena abbandonano il sacco ovigero. Invece, per la durata di un periodo noto come fase tollerante, i ragnetti si aggregano e molti collaborano nella costruzione d'una piccola tela. A volte attaccano persino piccole prede che incappano nella tela e avvolgono nella seta l'intruso. Dopo che parecchi giorni della fase tollerante sono passati, i ragnetti si disperdono, costruiscono ragnatele individuali e si cibano delle prede catturate. Tutti i ragnetti adottano simultaneamente il modello di comportamento solitario.

Va anche notato che in alcune specie di ragni solitari (tra cui membri delle famiglie degli eresidi, dei teridiidi e degli agelenidi) la femmina adulta non abbandona il sacco ovigero dopo averlo costruito, ma se lo porta appresso, finché ne emergono i ragnetti. A volte permette ai piccoli di condividere la preda catturata o li nutre con cibo rigurgitato o con speciali secrezioni. Questo tipo di cura parentale della prole ha una certa rassomiglianza con il livello inferiore della via evolutiva subsociale, secondo la terminologia di Michener, quindi persino tra le specie di ragni, riconosciute solitarie, esistono episodi transitori di socialità.

Quando i ragni vivono in gruppi, si verificano altri modelli di interazione. I gruppi si formano in una gran varietà di modi: per esempio i ragni adulti di alcune specie nelle famiglie degli uloboridi e degli araneidi si aggregano indipendentemente dal fatto che appartengano alla medesima prole di due genitori o a due figliolanze diverse. Ogni individuo in queste aggregazioni fila la propria tela. Tra alcune specie, gli individui possono anche fornire la propria seta come contributo alla ragnatela comune. Alcuni di questi gruppi possono essere costituiti persino da 1000 individui adulti. In generale ogni adulto vive indipendentemente, ma tutti condividono i vantaggi di una superficie d'una grande tela di gruppo e di un monopolio di habitat che potrebbe altrimenti esser conteso da specie competitive.

La vitalità delle aggregazioni semplici suddette dimostra l'esistenza nei ragni adulti singoli di un meccanismo di tolleranza, che deve essere abbastanza forte per lo meno da impedire ai ragni di mangiarsi a vicenda quando le prede sono scarse. Evidentemente tale meccanismo è anche specie-specifico, poiché non è limitato alla semplice garanzia che i ragni tollerino la presenza di tutti gli altri ragni della comunità: questi ragni sono tolleranti anche verso un qualsiasi loro simile appartenente alla medesima specie. Questo fatto è stato dimostrato nell'esperimento qui descritto. Alcuni individui della specie *Metepeira spinipes*, membro della famiglia degli araneidi, sono stati prelevati da popolazioni viventi a centinaia di chilometri di distanza e introdotti in aggregazioni locali di *M. spinipes*. La presenza di estranei non ha distrutto il meccanismo di tolleranza all'interno dell'aggregazione locale, e nemmeno si è notata alcuna differenza nel comportamento dei due gruppi.

Nei casi più avanzati di socialità dei ragni, esistono interazioni sostanzialmente più complesse di quelle che ho già descritto. Queste interazioni sono note solo per quattro (o forse cinque) specie di ragni. Due sono le specie africane: *Agelena consociata* e *Stegodyphus sarsinorum*. Gli altri sono ragni del nuovo mondo: *Anelosimus eximus* (e forse una seconda specie del genere, e cioè *A. studiosus*) nell'America del Sud, e una delle specie che ho potuto raccogliere in Messico, *Mallos gregalis*. Tutte queste specie hanno l'abitudine di costruire una grande ragnatela centrale che è occupata da tutti i ragni dell'aggregato e per generazioni successive.

Questi ragni collaborano anche nella cattura di prede assai più grosse di quelle che ciascuno di loro riuscirebbe a catturare da solo. Inoltre, dopo che la preda è stata catturata, i ragni la divorano tutti assieme. Le interazioni complesse che così si determinano richiedono che queste specie possiedano, oltre che un meccanismo di tolleranza, un'abilità nella coordinazione delle risposte individuali agli stimoli e la capacità di riconoscere indizi sensoriali intraspecifici o di reagire a qualche altro tipo d'informazione.

Bertrand Krafft dell'Università di Nancy ha osservato *Agelena consociata* nel Gabon e ha scoperto che la tolleranza a distanza ravvicinata in questa specie è mediata almeno in parte da indizi chemotattici. I membri non feriti della comunità si tollerano a vicenda, mentre un ragno ferito, o un ragno in cui l'odore normale superficiale sia stato artificialmente alterato mediante lavaggio in alcool o etere, viene attaccato immediatamente. Gli indizi chemotattici, come pure altri fattori del meccanismo di tolleranza di questi ragni, probabili ma non ancora identificati, non sono affatto limitati a popolazioni locali di questa specie: come per *Matepeira spinipes*, i ragni singoli della medesima specie possono essere spostati da una colonia all'altra senza alterare il tipico modello di attività comunitaria.

Non esistono prove per affermare che una qualsiasi di queste specie di ragni abbia evoluto un sistema di caste: ossia gli adulti non hanno forme diverse a seconda della divisione del lavoro. Alcune differenze nel ruolo di comportamento si possono presentare per effetto dell'età o per variazione dei ritmi biologici, ma non si è ancora scoperto quali siano gli stimoli indicatori forniti dai singoli ragni per ottenere la cooperazione necessaria. Questo tipo di comportamento è tuttavia un esempio di socialità che non viene facilmente uguagliata da alcuna specie sociale d'insetto, e potrebbe essere ascritto a una speciale categoria: comportamento comunitario-cooperativo.

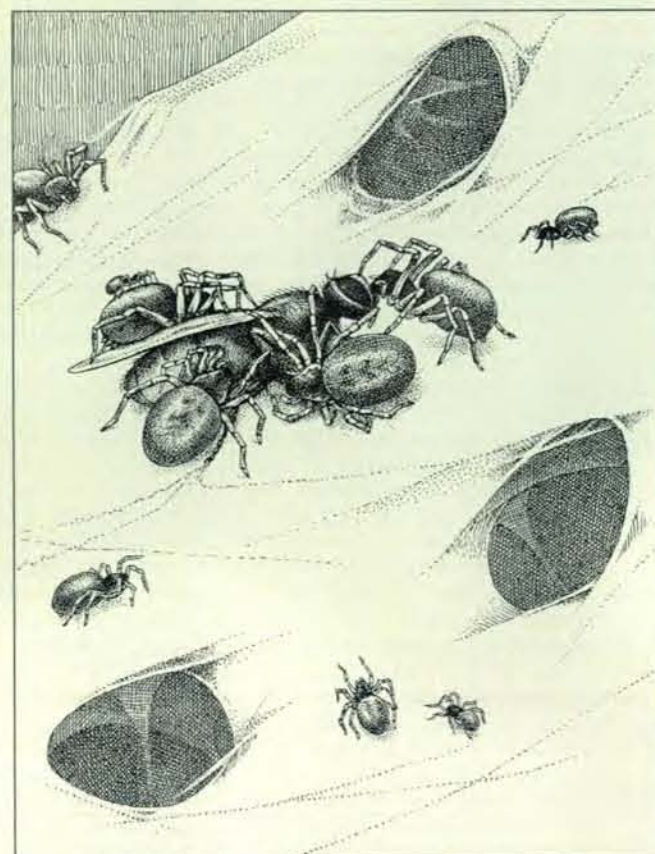
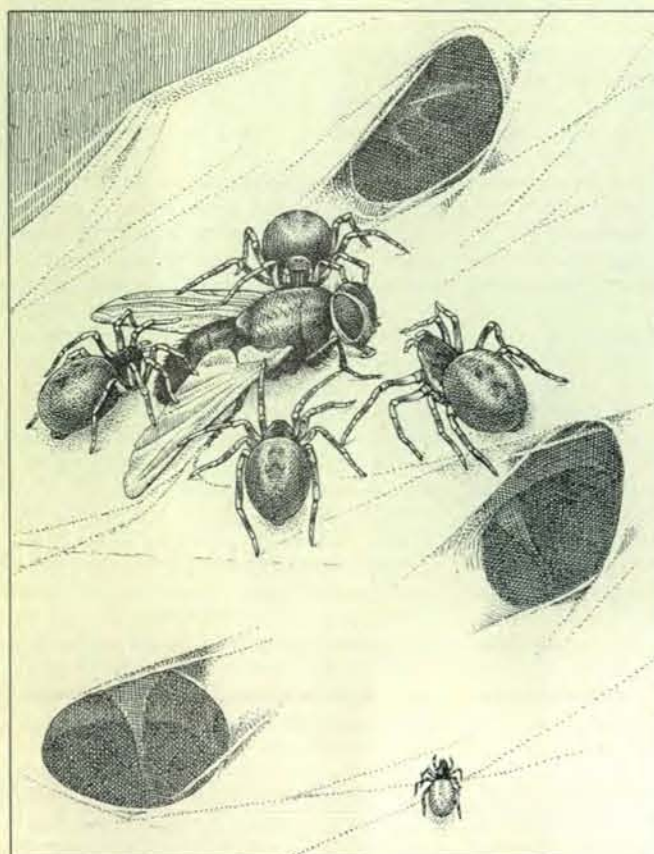
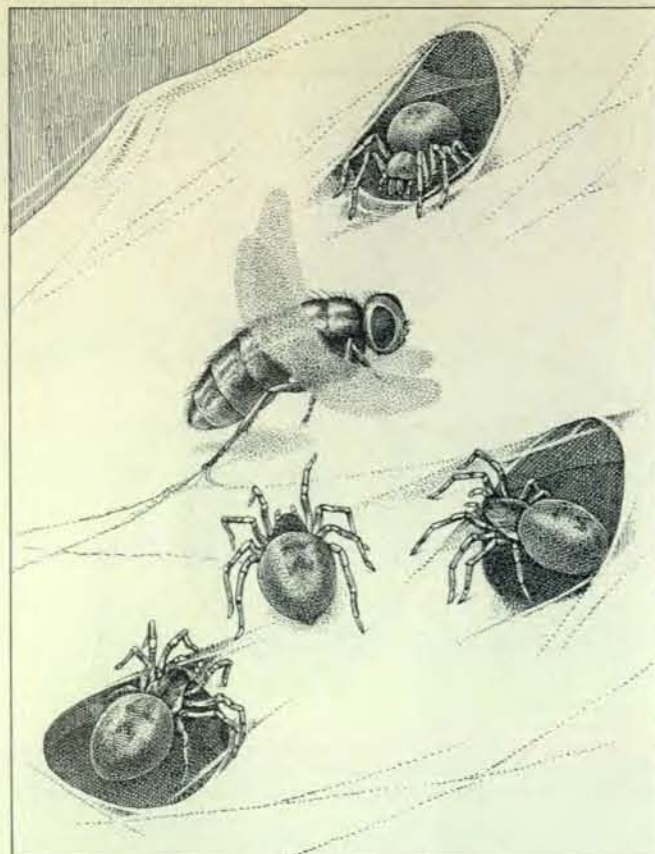
Il ragno sociale messicano *Mallos gregalis* intrappola soprattutto mosche sulla superficie appiccicosa della ragnatela comune a forma di lenzuolo, che viene filata attorno al ramo d'un albero. I messicani lo hanno denominato *el mosquero*, ossia l'ammazza-mosche, e nella stagione delle piogge, quando le mosche domestiche sono particolarmente fastidiose, i contadini che vivono nei pressi di Guadalajara portano un ramo ricoperto da questa ragnatela in casa, nel medesimo modo in cui altri appendono la carta moschicida. Membro della famiglia dei dictinidi, *Mallos gregalis* è un ragno cribellato: possiede una piastra piena di forellini, un crivello appunto, nella superficie inferiore dell'addome (*si veda l'illustrazione nella pagina a fronte*). La seta appiccicosa emerge dai fori finissimi del crivello e viene pettinata mediante le due zampe posteriori, dotate d'una fila speciale di setole: quest'organo viene denominato calamistro.

È la seta che forma le aree appiccicose, che funzionano da trappola per la preda, sulla parte esterna della ragnatela. La tela, nel suo complesso, è una struttura elaborata che comprende fili di sostegno tra la superficie principale, i rami e le foglie, zone riparate per i ragni e speciali camere dove le femmine vivono con i loro sacchi ovigeri, involucri fini di seta che contengono da 10 a 20 uova.

La tela comune di *M. gregalis* può raggiungere enormi dimensioni. Ne ho potuto osservare una presso presso Guadalajara che copriva i rami nei tre quarti superiori di un albero della famiglia delle mimose, alto circa 20 metri. All'attacco dei rami sul tronco la seta della tela era grigia, ma presso le punte era nuova e bianca: evidentemente la costruzione stava progredendo verso l'esterno. I ragni erano attivi su tutte la parti della ragnatela.

Le osservazioni condotte sia in natura, sia in laboratorio, confermano che la costruzione della ragnatela di *M. gregalis* viene compiuta con un lavoro di tutti i componenti della colonia. Se una colonia di ragni tenuta in laboratorio ha a sua disposizione un supporto in qualche modo paragonabile a un albero, come un bastone verticale, costruisce la caratteristica ragnatela avviluppante; in assenza di tale supporto, i ragni costruiscono un modello di ragnatela tridimensionale, tipico di altre specie della famiglia dei dictinidi. Benché questa tela abbia un aspetto diverso da quella naturale, comprende camere di ritirata e d'incubazione per i sacchi ovigeri. Nella ragnatela del laboratorio un compito iniziato da un ragno può esser finito da un altro. Ho anche visto un ragno della colonia filare trefoli di seta ordinaria, dopo di che altri ragni aggiungevano fasce di seta appiccicosa prodotta dal crivello.

Osservati in natura, i ragni sembrano muoversi attorno a caso e senza fretta, entrando e uscendo da buchi situati nella superficie della tela. Le mosche vengono catturate dalla tela appiccicosa: quando una mosca rimane impastoiata, due o tre ragni si avvicinano all'insetto che ronza disperatamente dibattendosi, lo immobilizzano coi loro morsi velenosi e poi lo divorano.

Il comportamento predatorio dei ragni può essere osservato in dettaglio in labo-

In questi disegni è ricostruita, sulla base delle osservazioni condotte in laboratorio, la reazione prodotta in una colonia di ragni comunitari da parte di un intruso. A differenza dei membri d'una colonia selvatica, tutti i ragni della colonia di laboratorio hanno digiunato per lo stesso periodo di tempo. Il suono d'una mosca che passa è percepibile dall'orecchio d'un uomo, ma non attira l'attenzione dei ragni. Anche quando una mosca si posa sulla superficie della tela (*in alto a sinistra*), solo i ragni vicini cambiano direzione. Ma il ronzio disperato d'una mosca invischiata nella seta appiccicosa (*in alto a destra*) stimola una risposta che si estende a tutta la colonia, e i ragni avanzano verso la preda a saltelli veloci. I primi morsi (*in basso a sinistra*) provocano da parte della mosca un ronzio ancor più forte, che stimola altri ragni ad avvicinarsi. Quando il festino comincia (*in basso a destra*) anche i ragnetti immaturi raggiungono gli adulti.

ratorio. Noi diamo da mangiare alle nostre colonie una volta ogni 5 giorni. Grazie a questa tecnica esiste una maggiore probabilità che i ragni si apprestino a cibarsi contemporaneamente. In qualsiasi momento uno o due individui di una colonia di 100 ragni sono di solito sulla superficie della tela, mentre gli altri si trovano all'interno. Quando una mosca domestica viene posta nella gabbia dei ragni e vola attorno, produce un ronzio percettibile allo sperimentatore, ma non provoca alcun cambiamento apparente nell'attività casuale dei ragni.

Una mosca che si posa su una parte non appiccicosa della ragnatela e vi passeggia stimola una risposta localizzata: alcuni ragni si dirigono verso la posizione occupata dalla mosca, ma limitano le loro reazioni solo a questo cambiamento di direzione. Se però la mosca rimane impigliata nella parte appiccicosa della tela e comincia a ronzare all'impazzata, il comportamento dei ragni muta improvvisamente: da ogni parte della ragnatela i ragni che sono rimasti a riposo si volgono verso la mosca intrappolata e iniziano ad avvicinarsi, compiendo brevi saltelli. La mosca continua a ronzare anche dopo che i primi ragni, giunti a ridosso, cominciano il loro attacco, di solito mordendole una zampa o un'ala. Il ronzio insistente attira altri attaccanti, che si dirigono verso la preda che alla fine scompare quasi del tutto sotto la massa di ragni affamati: attaccano sia i ragni maschi sia le femmine e persino i ragni immaturi prendono parte al festino, brulicando sopra gli adulti in cerca d'un posto adatto per cibarsi.

Benché i ragni attaccanti in una colonia tenuta in gabbia siano particolarmente aggressivi, non abbiamo mai notato un ragno che ne aggrediva un altro. Poiché prendiamo in considerazione tutto il repertorio di comportamento che differenzia i ragni sociali da quelli solitari, questo aspetto del nutrirsi in aggregazione è significativo: per esempio i ragni giovani solitari della specie *Araneus diadematus*, quando sono artificialmente confinati in zone ravvicinate, cominciano a cercar cibo in comune. Tuttavia tra i ragnetti solitari artificialmente confinati, un meccanismo di tolleranza, anche qualora esistesse, opera solo in maniera imperfetta. I ragnetti infatti in questo caso non solo divorano le mosche catturate, ma si divorano anche vicendevolmente. Da ciò si deduce che esiste un forte meccanismo di tolleranza che spiega la nutrizione comunitaria in *Mallos gregalis*, così come esiste un meccanismo di coordinazione che spiega la collaborazione nella cattura della preda.

Il meccanismo di tolleranza che esiste nelle colonie di *M. gregalis* viene ora studiato nei nostri laboratori. Dalle osservazioni compiute risulta evidente che si tratta di un meccanismo potente e che opera sia a distanza ravvicinata sia a grande distanza. Anzi, per meglio dire, potrebbero essere in funzione parecchi meccanismi separati, forse mediati da sistemi indicatori che permettono la distinzione, per esempio, tra le vibrazioni

I rami ricoperti di ragnatela d'una specie di mimosa, nei pressi di Guadalajara, fanno da supporto alla ragnatela comunitaria d'una colonia di *Mallos gregalis*. I buchi sparsi permettono ai ragni di spostarsi liberamente dalle zone interne della ragnatela alla superficie appiccicosa esterna, dove gli intrusi possono restare intrappolati. Nella stagione in cui abbondano le mosche, la gente del luogo porta spesso questi rami in casa, per usarli come «carta moschicida».

delle ragnatele provocate da una preda e quelle provocate da membri della colonia. Per dimostrare la fondatezza di questa ipotesi, stiamo sottoponendo le colonie agli stimoli di svariate vibrazioni della ragnatela, nella speranza di isolare queste indicazioni.

Il comportamento sociale della seconda specie di ragno messicano, *Oecobius civitas*, dapprima sembra essere principalmente aggregativo, come il comportamento di altri ragni che costruiscono i nidi in stretta vicinanza. L'oscurità del suo microhabitat rende difficili le osservazioni sul comportamento, tuttavia si è potuto scoprire il metodo insolito per la cattura della preda. *O. civitas* possiede un organo a forma di dito, il tubercolo anale, posto sull'addome presso le filiere che secernono i fili di seta: con questa appendice può filare la seta appiccicosa al di fuori del suo crivello a mo' di fune, avvolgendola attorno alla preda (*si veda l'illustrazione a pagina 89*).

Studiando più da vicino la socialità di *O. civitas* si può dimostrare che questa specie si aggrega in modo piuttosto complicato. Il comportamento del ragno mette in evidenza una curiosa combinazione di atteggiamenti di tolleranza e di fuga. Sulla faccia inferiore della roccia che ripara i ragni, ogni individuo fila un piccolo tubo di seta, con un'estremità aperta, che costituisce il nascondiglio; attorno a questo tubo costruisce una tela sottile, vicino alla superficie della roccia, che costituisce un sistema d'allarme. La ragnatela è quindi costituita da queste due strutture, e si adatta a una cavità o a una spaccatura della roccia. Se un ragno viene disturbato e scacciato dal suo nascondiglio, attraversa velocissimo la superficie della roccia e se manca una spaccatura libera nella roccia per potervisi nascondere, può ricercare un rifugio nel nascondiglio d'un altro ragno della medesima specie. Se l'altro ragno è in casa quando l'intruso vi penetra, non lo attacca, ma scappa e cerca a sua volta un nuovo rifugio. Così non appena il primo ragno è disturbato, il processo di spostamento può a volte continuare per parecchi secondi, provocando per la maggior parte dei ragni dell'aggregato uno slittamento dal proprio rifugio a un rifugio estraneo.

Mediante osservazioni compiute nell'ambiente naturale ed esperimenti di laboratorio, si è potuto dimostrare che, come in *Metepeira* e *Mallos*, i meccanismi implicati in questo strano miscuglio di tolleranza e di fuga si estendono al di là della popolazione locale, e interessano altri ragni della medesima specie. Inoltre, all'interno della popolazione locale lo slittamento verso il rifugio di un altro ragno può essere un processo semipermanente: quando i ragni rimangono indisturbati, occupano una posizione fissa

nella ragnatela per lunghi periodi. In ogni evenienza, il modello di comportamento di questa specie avvantaggia il singolo ragno, poiché gli fornisce più d'una possibilità di ritirata.

Il comportamento di gruppo di *Oecobius civitas*, di gran lunga più semplice rispetto a quello di *Mallos gregalis*, è tuttavia molto efficace, poiché permette ai ragni di vivere assieme in condizioni d'affollamento. Indubbiamente il meccanismo di fuga dà un contributo alla capacità dei ragni di mantenere una densità elevata nel caratteristico microhabitat ristretto. Tra gli altri fattori si annoverano probabilmente l'insolita tecnica predatoria del ragno e la spaziatura delle ragnatele individuali. I meccanismi che stanno alla base della fuga e della tolleranza non sono stati ancora chiariti, ma certamente costituiscono gli elementi basilari d'un comportamento di gruppo più complesso.

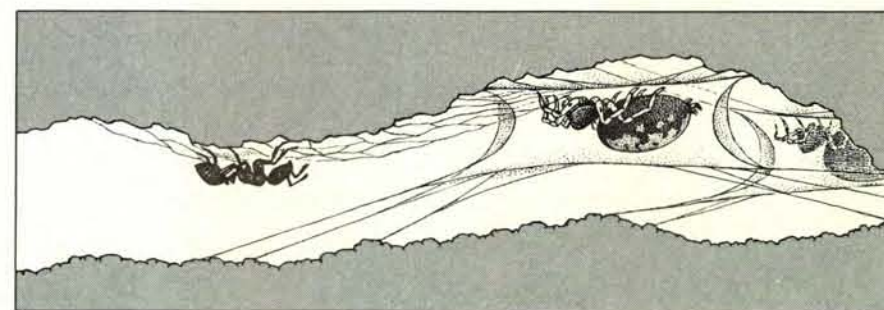
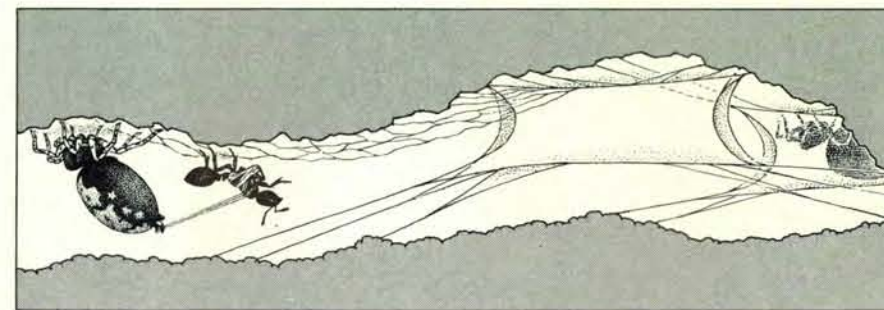
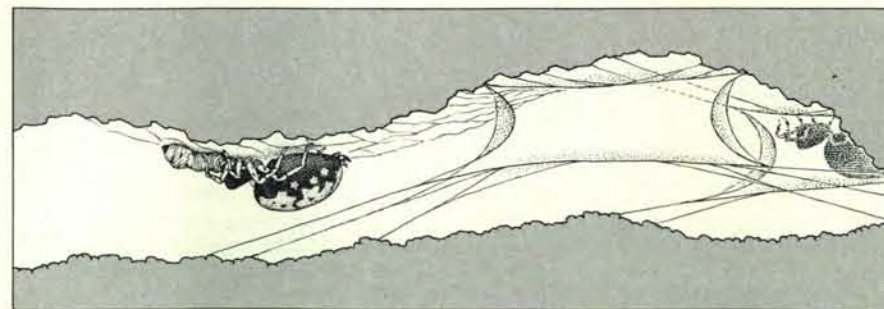
Si è fatto notare che *Oecobius civitas* dimostra un tipo di socialità ancor più notevole: la costruzione d'un sacco ovigero comunitario da parte delle femmine del gruppo. La possibilità fornita da tale progresso nel campo del comportamento è venuta in luce recentemente, quando William A. Shear dello Hampton-Sydney College ha intrapreso una ricerca sistematica sui ragni della famiglia degli ecobidi. In questo lavoro è stato aiutato da alcuni colleghi, che gli hanno donato esemplari per completare il suo progetto. Tra i donatori vi è stato Willis J. Gertsch, dell'American Museum of Natural History, che aveva raccolto esemplari di *O. civitas*, la sua ragnatela e i sacchi ovigeri nella zona di Guadalajara.

Il sacco ovigero normale degli ecobidi contiene da 5 a 10 uova. Nel materiale donato da Gertsch, Shear trovò tuttavia due gruppi di più di 200 ragni immaturi. Ogni gruppo era contenuto in una struttura che aveva tutto l'aspetto d'un singolo sacco ovigero. Shear pubblicò le sue osservazioni nel 1970, proponendo l'ipotesi che *O. civitas* potesse deporre uova in un sacco ovigero comunitario.

Quando ho catturato esemplari di *O. civitas*, con i sacchi ovigeri, in una zona presso le coste del lago Sayula, dove Gertsch aveva compiuto questa raccolta, ho trovato che parecchie altre specie di ragni condividevano l'habitat roccioso con gli ecobidi. Ho potuto perciò raccogliere una notevole varietà di sacchi ovigeri. Ho poi sigillato i singoli sacchi, ciascuno in un tubo diverso. Ho avuto la sorpresa di trovare che solo i sacchi piccoli, contenenti in media sette uova ciascuno e raccolti nella ragnatela di *O. civitas* o nei pressi, schiudendosi fornivano ragnetti di ecobidi.

Dopo aver allevato questo ragno in laboratorio per tre generazioni e aver osservato unicamente sacchi contenenti da 5 a 10 uova, ritengo che questo sia il modello normale di comportamento riproduttivo di *O. civitas*.

L'accoppiamento tra i ragni non è stato ancora osservato nelle nostre popolazioni di laboratorio di *Mallos gregalis* e *Oecobius civitas*. I ragni maschi solitari attraversano fasi in cui elaborano prima dell'accoppiamento complesse manovre, i cosiddetti schemi di corteggiamento, che forse inibiscono la tendenza alla cattura predatoria da parte della femmina durante la copula. Tra i ragni sociali, che vivono in aggregazioni tolleranti, tali manovre non sembrerebbero necessarie. In effetti, se esistono veramente delle differenze nel comportamento di accoppiamento tra i ragni solitari e quelli sociali, queste differenze potrebbero fornire indicazioni per interpretare l'evoluzione della socialità nei ragni. A questo riguardo abbiamo fatto un'osservazione probabilmente significativa intorno alla fecondità: i ragni solitari allevati in laboratorio mantengono i ritmi ciclici di riproduzione caratteristici dello stato naturale; ma quando le nostre colonie di *M. gregalis* vengono immesse in un ambiente uniforme e in periodi controllati di buio e di luce, producono uova feconde durante il corso dell'anno.

Le osservazioni compiute sui due ragni messicani hanno fornito una notevole messe di dati circa la loro socialità, ma non hanno ancora permesso di risolvere tutti i quesiti. Non sappiamo, per esempio, quali condizioni favoriscano lo sviluppo della socialità tra i ragni e nemmeno quali meccanismi siano coinvolti nella tolleranza, nella tendenza alla fuga, nella formazione dei gruppi o nella coordinazione dell'attività. Inoltre non sappiamo quali rapporti di parentela abbiano le diverse forme di socialità tra i ragni e come venga trasportata all'interno del gruppo l'informazione, durante le interazioni complesse. La ricerca di risposte tuttavia sembra offrire una certezza: più dati raccoglieremo intorno alla socialità degli animali relativamente semplici, più sapremo comprendere la socialità delle specie più complesse, ivi compresa quella umana.

La cattura d'una preda, di solito una formica in cerca di cibo, da parte di *Oecobius civitas*, segue un modello complicato che comincia quando l'intrusa disturba la rete d'allarme del ragno.

Il ragno, avvisato abbandona il rifugio e si muove in cerchi attorno alla preda, con l'addome in avanti e sollevato, mentre pettina una striscia di seta appiccicosa con il tubercolo anale.

Avvolta nella seta, la formica rimane immobilizzata. Il ragno può a volte riposarsi per un certo tempo oppure può voltarsi (*a sinistra*) per morderla e paralizzarla completamente. Solo il ragno che ha immobilizzato la preda la divorerà; gli altri ragni anche vicini non partecipano al festino.

GIOCHI MATEMATICI

di Martin Gardner

Recenti progressi nel campo dei quadrati e dei cubi magici

«Durante la mia gioventù, nei momenti d'ozio (che probabilmente avrei potuto impiegare in modo più costruttivo), mi divertii a costruire... quadrati magici.»

—Da una lettera di Benjamin Franklin

Lo studio dei quadrati e dei cubi magici ha registrato recentemente due importanti progressi: l'enumerazione di tutti i quadrati magici di ordine 5 e la costruzione del primo cubo magico perfetto. Sono lieto di essere il primo a dare il giusto risalto in questa sede a questi due risultati, e perché si possa più completamente realizzarne la portata traccerò a grandi linee la storia dei quadrati magici.

Sebbene alcuni dei più grandi matematici si siano occupati di quadrati magici e sebbene questi lavori abbiano avuto attinenza con settori certamente non banali della matematica, come la teoria dei gruppi, i reticoli, i quadrati latini, i determinanti, le partizioni, le matrici, le relazioni di congruenza in aritmetica, i cultori più entusiastici di questo campo sono stati i dilettanti. Il famoso quadrato di Franklin, una matrice di 16 per 16 ingegnosamente costruita che secondo Benjamin Franklin era «il più magicamente magico tra tutti i quadrati magici mai costruiti da un mago», costituisce l'oggetto di vari articoli e monografie. Sui quadrati magici è stato scritto molto, soprattutto da parte di persone che pur non occupandosi professionalmente di matematica erano rimaste invischiate nelle eleganti simmetrie di questi schemi intrecciati di numeri.

Un quadrato magico standard, come quasi tutti i lettori sapranno, è una disposizione quadrata di numeri interi positivi da 1 a N^2 tale che la somma di ogni riga, ogni colonna e ognuna delle due diagonali è costante. Il numero N è l'ordine del quadrato. Si vede facilmente che la costante magica è data dalla somma di tutti i numeri costituenti il quadrato divisa per N; la formula che la fornisce è $(1+2+3...+N^2)/N = \frac{1}{2}(N^3+N)$.

Quando l'ordine è 1 abbiamo il caso banale del quadrato costituito dal numero 1, che ovviamente è unico. Si dimostra facilmente che non esistono quadrati magici di ordine 2.

Vi sono otto modi di disporre le cifre da 1 a 9 in una matrice di ordine 3 che sia un quadrato magico. Se come è d'uso non si contano le rotazioni e le riflessioni, risulta che il quadrato magico di ordine 3 è unico. Per apprezzare in tutto il suo fascino la bellezza di questa che è la più antica di tutte le curiosità combinatorie, si considerino i modi in cui la sua costante 15 può esprimersi come somma di triple di numeri positivi diversi. I modi distinti sono esattamente otto:

9+5+1
9+4+2
8+6+1
8+5+2
8+4+3
7+6+2
7+5+3
6+5+4

Ora nel quadrato di ordine 3 vi sono otto linee che devono dare come somma 15: le sei ortogonali (righe e colonne) e le due diagonali. Queste otto linee corrispondono esattamente alle otto triple di numeri precedenti. Dato che il numero centrale del quadrato deve appartenere alle due diagonali, a una riga e a una colonna, deve necessariamente trattarsi di una cifra che compare in quattro delle otto triple. L'unica cifra con queste caratteristiche è il 5 che sarà quindi il numero centrale.

Consideriamo la cifra 9. Essa appartiene solo a due triple e quindi non può essere collocata in un angolo, perché le cifre d'angolo appartengono a tre linee. Si tratterà quindi di una cifra laterale. Data la simmetria del quadrato, non importa in quale casella laterale si collochi il 9, scegliamo quindi in modo arbitrario quella sopra al 5. Negli angoli superiori, ai due lati del 9, non possiamo collocare altro se non il 2 e il 4; anche in questo caso la simmetria del quadrato fa sì che i due casi possibili siano immagini speculari l'uno dell'altro. Il resto del quadrato segue in modo automatico; con questa semplice costruzione abbiamo dimostrato l'unicità del quadrato.

Il quadrato magico di ordine 3, nella forma mostrata nell'illustrazione in alto di pagina 96, è il *Lu shu* dell'antica civiltà cinese. Secondo la leggenda lo schema del quadrato fu rivelato per la prima volta all'uomo dal guscio di una tartaruga sacra emersa dalle acque del fiume Lo nel XXIII secolo avanti Cristo; secondo gli studiosi cinesi contemporanei lo schema non è anteriore al IV secolo avanti Cristo. Da allora fino al X secolo questo quadrato magico è stato uno dei simboli mistici cinesi più importanti. I numeri pari rappresentavano il principio femminile dello yin, i dispari quello maschile dello yang, il numero 5 al centro rappresentava la terra e i numeri circostanti i quattro elementi, con una equilibrata presenza di yin e di yang: 4 e 9 rappresentavano i metalli, 2 e 7 il fuoco, 1 e 6 l'acqua, 3 e 8 il legno. Per altre notizie storiche sul *Lu shu* e per le sue connessioni con l'arte divinatoria e con l'*I Ching* si vedano gli articoli *The Magic Square of Three in Old Chinese Philosophy and Religion*, di Schuyler Camman, in «History of Religions», vol. 1, estate 1961, e *Old Chinese Magic Squares*, dello stesso autore, in «Sinologica», vol. 7, 1962.

I quadrati magici di ordine 4 sono 880 (se si escludono rotazioni e riflessioni) e il primo a enumerarli è stato Bernard Frénicle de Bessy nel 1693. Esistono diversi modi di classificarli, uno dei migliori è dovuto a Henry Ernest Dudeney, che spiega il suo metodo in un ottimo articolo sui quadrati magici pubblicato nelle prime ristampe della quattordicesima edizione dell'*Enciclopedia Britannica*. L'ultima ristampa di quell'edizione riporta al posto dell'articolo di Dudeney un ottimo articolo storico di Camman. Nell'edizione attuale (la quindicesima) c'è solamente un banale articolo sui quadrati magici nella *Micropaedia*.

Quanti sono i quadrati magici di ordine 5? Albert Candy, nel suo *Construction, Classification and Census of Magic Squares of Order Five*, pubblicato a sue spese a Lincoln, Nebraska, nel 1938, è stato quello che più si è avvicinato al numero esatto stimandoli in numero di 13 288 952. La cifra esatta si è conosciuta solo nel 1973, quando l'enumerazione fu completata grazie a un programma di calcolo sviluppato da Richard Schroeppel, un matematico programmatore della Information International. Il programma si avvale di una procedura back-tracking standard, consiste di circa 3500 parole e richiede circa 100 ore di tempo di esecuzione su un PDP-10. Nell'ottobre dello scorso anno è stato anche pubblicato un resoconto finale sulla questione, a opera di Michael Beeler.

La stima di Candy era di molto in difetto. Se non si contano rotazioni e riflessioni, i quadrati magici di ordine 5 sono 275 305 224, ma Schroeppel preferisce dividere questo numero per 4 ottenendo in totale 68 826 306 quadrati magici. La ragione di questa divisione sta nel fatto che oltre alle otto varianti che si ottengono per mezzo di rotazioni e riflessioni ve ne sono altre quattro generate dalle due seguenti trasformazioni che conservano il carattere magico del quadrato:

1. Si scambino tra loro le colonne che costituiscono il bordo destro e quello sinistro, e così pure le righe che costituiscono il bordo superiore e quello inferiore.

2. Si scambino tra loro le righe 1 e 2 e le righe 4 e 5, e così pure le colonne 1 e 2, e le colonne 4 e 5.

Combinando queste due trasformazioni con le due riflessioni e le quattro rotazioni si ottengono in totale 2×4×2×2=32 forme che si possono ritenere isomorfe. Con questa definizione di isomorfismo il loro numero risulta appunto 68 826 306.

Tale numero può essere ulteriormente abbassato tenendo conto di un'altra trasformazione ben nota. Se a ogni numero di un quadrato magico si sostituisce la sua differenza da N^2+1 (nel nostro caso 26), il quadrato risultante viene detto complemento ed è ancora magico. Quando il centro di un quadrato magico di ordine 5 è 13, esso è isomorfo al suo complemento, mentre in caso contrario non è isomorfo. Se si estende il concetto di quadrato isomorfo fino a includere i complementi, il numero dei quadrati magici di ordine 5 scende a circa 35 milioni.

La classificazione dei quadrati di ordine cinque in categorie significative è una impresa non da poco. Dudeney scrisse un tempo che certi criteri di classificazione in tipi dei quadrati magici gli sembravano utili quanto distinguere la gente nelle due categorie di chi fiuta tabacco e di chi non lo fa. Ciononostante alcune di queste classificazioni conducono a risultati inaspettati. Si consideri per esempio il numero totale dei quadrati di ordine 5 che hanno come centro i numeri da uno a 13:

1. 1 091 448
2. 1 366 179
3. 1 914 984
4. 1 958 837
5. 2 431 806
6. 2 600 879
7. 3 016 881
8. 3 112 161
9. 3 472 540
10. 3 344 034
11. 3 933 818
12. 3 784 618
13. 4 769 936

Si noti che il numero dei quadrati aumenta rapidamente passando da 1 a 8, mentre diminuisce passando da 9 a 10 e da 11 a 12. Sorprendente è la constatazione che vi sono più quadrati con centro 11 che quadrati con centro 12, e così pure per il caso di 9 e di 10. Ovviamente la stessa singolarità si riscontra per i quadrati i cui centri vanno da 14 a 25, dato che ogni quadrato con centro diverso da 13 ha un complemento (non isomorfo). I quadrati con centro 1 sono tanti quanti quelli con centro 25, e così per gli altri numeri tranne il 13. I lettori interessati ai particolari del programma di Schroeppel possono scrivergli a questo indirizzo: 835 Ashland Avenue, Santa Monica, California 90405.

L'illustrazione in basso mostra un quadrato di ordine 5 che è, per così dire, più magico di ogni altro. È associativo, il che significa che ogni coppia di numeri simmetricamente opposta rispetto al centro dà come somma N^2+1, e pandiangonale (talora detto anche diabolico), il che significa che le sue diagonali spezzate danno come somma la costante magica 65. Vale a dire che se tasselliamo il piano con questo quadrato possiamo isolare ovunque in questo schema un quadrato

1	15	24	8	17
23	7	16	5	14
20	4	13	22	6
12	21	10	19	3
9	18	2	11	25

∞

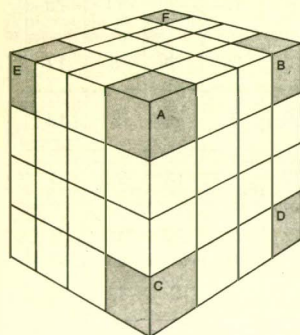
1	15	24	8	17	1	15	24	8	17	1	15	24	8	17
23	7	16	5	14	23	7	16	5	14	23	7	16	5	14
20	4	13	22	6	20	4	13	22	6	20	4	13	22	6
12	21	10	19	3	12	21	10	19	3	12	21	10	19	3
9	18	2	11	25	9	18	2	11	25	9	18	2	11	25
1	15	24	8	17	1	15	24	8	17	1	15	24	8	17
23	7	16	5	14	23	7	16	5	14	23	7	16	5	14
20	4	13	22	6	20	4	13	22	6	20	4	13	22	6
12	21	10	19	3	12	21	10	19	3	12	21	10	19	3
9	18	2	11	25	9	18	2	11	25	9	18	2	11	25

∞ ∞ ∞

Un quadrato magico associativo e pandiagonale di ordine 5 (in alto) e le sue permutazioni cicliche (in basso).

Il quadrato magico Lo shu *dell'antica Cina.*

di 5 per 5 essendo certi che sarà magico sebbene non necessariamente associativo. Per ottenere un quadrato che sia anche associativo, bisogna che il numero 13 sia collocato nel centro.

Il *Lu shu* è associativo ma non pandiagonale. Un quadrato di ordine 4 può essere associativo o pandiagonale («o» esclusivo). Il quadrato di ordine 5 è il più piccolo quadrato in cui si possono assommare queste due caratteristiche. Se si escludono le rotazioni e le riflessioni, i quadrati pandiagonali di ordine 5 sono 3600, numero che si riduce a 144 se escludiamo anche le varianti ottenute con la permutazione ciclica di righe e di colonne. In altre parole, esistono 144 schemi infiniti del tipo qui mostrato, contenenti ognuno 25 quadrati pandiagonali di ordine 5. Di questi 144 solo 16 contengono un quadrato che è anche associativo. Tutto ciò, detto tra parentesi, era noto prima che Schroeppel sviluppasse il suo programma.

Dei sedici quadrati associativi e pandiagonali di ordine 5, ve ne sono quattro con 1 nella prima casella, quattro con 1 nella terza, quattro con 1 nella settima e quattro con 1 nell'ottava. Durante il medioevo i musulmani attribuivano un valore particolare ai quadrati pandiagonali con la cifra 1 nel centro. Gli schemi non erano ovviamente associativi, ma la cifra 1 collocata al centro simboleggiava l'unità di Allah. Il loro atteggiamento reverente nei confronti di questo simbolo arrivava al punto di lasciare vuota la casella centrale.

È possibile estendere in modo del tutto naturale il concetto di quadrato magico al caso di tre o più dimensioni. Un cubo magico perfetto è una disposizione cubica dei numeri interi positivi da 1 a N^3 tale che prese comunque N caselle allineate la loro somma sia costante. Tali linee includono le ortogonali (le linee parallele a uno spigolo), le due diagonali principali di ogni sezione trasversale ortogonale e le quattro diagonali spaziali. La costante è $(1+2+3+\ldots+N^3)/N^2 = \frac{1}{2}(N^4+N)$.

L'ordine 1 ha un solo cubo perfetto banale, e si dimostra facilmente che l'ordine 2 non ne ha. Per quanto riguarda l'ordine 3, anch'esso non ammette cubi perfetti. Non so chi sia stato il primo a dimostrarlo, ma Richard Lewis Myers Jr. ha fornito una dimostrazione particolarmente semplice. Si consideri una qualsiasi sezione trasversale di 3 per 3. Supponiamo che A, B, C siano i numeri della prima riga, D, E, F i numeri della terza e X il numero centrale. Dato che le diagonali e la colonna centrale devono dare tutte come somma 42, possiamo scrivere:

$$3X+A+B+C+D+E+F=3(42).$$

Da questo numero sottraiamo $A+B+C+D+E+F=2(42)$ ottenendo $3X=42$ da cui $X=14$. Dato che il numero 14 non può essere il centro di ogni sezione trasversale, il cubo di ordine 3 risulta essere impossibile.

Contrariati dalla mancata esistenza di questo cubo, i cultori di cubi magici hanno indebolito i requisiti definendo una specie di cubo magico semiperfetto che esiste in tutti gli ordini maggiori di 2. Chiameremo questi cubi, in cui sono magiche solo le ortogonali e le quattro diagonali spaziali, cubi di Andrews, dal nome di W.S. Andrews, autore di un libro *Magic Squares and Cubes* (1917) (ristampato nei Dover paperback) in cui dedica due capitoli a questo tipo di cubi. Il cubo di Andrews di ordine 3 deve essere associativo e deve avere il numero 14 al centro. John R. Hendricks ha dimostrato («Journal of Recreational Mathematics», vol. 5, gennaio, 1972, pagg. 43-50) che esistono quattro cubi di questo tipo, a meno di rotazioni e riflessioni. Andrews li fornisce tutti anche se non sembra essersi accorto che esauriscono tutti i tipi fondamentali possibili.

Non esistono cubi perfetti di ordine 4. Per quanto ne so la prima dimostrazione di questo fatto fu data da Schroeppel in un memorandum del 1972. Il primo passo consiste nel dimostrare che in una qualsiasi sezione di 4 per 4 (ortogonale o diagonale) i quattro angoli devono dare sempre come somma la costante. Supponiamo che Q sia la costante e indichiamo con lettere diverse le 16 caselle (*si veda la figura al centro di questa pagina*). Le linee colorate indicano sei quadruple che coprono le 16 caselle. Dato che ogni casella d'angolo è comune a tre linee, $3A+3D+3M+3P$ più ognuna delle altre caselle presa una sola volta dà come somma $6Q$. Se da questo numero sottraiamo i valori delle quattro righe restiamo con $2A+2D+2M+2P=2Q$ che si riduce a $A+D+M+P=Q$. Questo è il nostro primo lemma.

Consideriamo ora gli otto angoli del cubo. Dimostreremo che due angoli qualsiasi connessi da uno spigolo devono dare come somma $Q/2$. Chiamiamo A e B i due angoli e C, D e E, F gli angoli corrispondenti a due qualsiasi spigoli paralleli allo spigolo che connette A e B (*si veda la figura in basso di questa pagina*). $ABCD$, $EFBA$, $EFDC$ sono gli angoli di sezioni trasversali di 4 per 4, quindi la loro somma è $3Q$. Raccogliendo i termini comuni:

$$2A+2B+2C+2D+2E+2F=3Q.$$

Dividendo ambo i membri per 2 otteniamo:

$$A+B+C+D+E+F=3Q/2.$$

Da ciò sottraiamo $C+D+E+F=Q$ per ottenere $A+B=Q/2$. Questo è il nostro secondo lemma.

Consideriamo ora l'angolo B, connesso agli angoli A, D, F. Dato che $A+B=F+B=D+B$, possiamo sottrarre B da ogni membro, ottenendo $A=F=D$. Ciò tuttavia è assurdo, il che conclude la nostra dimostrazione. Esiste un cubo magico perfetto di ordine 5? È un problema irrisolto. Schroeppel ha fatto un primo passo dimostrando (valendosi di mezzi algebrici e combinatori) che se un tale cubo esiste, il suo centro deve essere 63. Non si sa se esistano cubi magici negli ordini 6 e 7.

Esistono cubi magici perfetti di ordine 8. Un metodo di costruzione è stato scoperto nella primavera del 1970 da Myers, quando aveva solo 16 anni ed era studente presso la William Tennant High School a Johnsville, Pennsylvania. A quei tempi egli mi spedì una breve nota sulla questione, dicendo che aveva costruito il suo primo cubo «dopo tre mesi, sette teorie e 31 fogli di carta millimetrata». Devo ammettere che in un primo tempo non ho dato importanza alla cosa. Myers non mi aveva spedito un cubo ef-

La prova di Schroeppel, lemma 1.

La prova di Schroeppel, lemma 2.

1

19	497	255	285	432	78	324	162
303	205	451	33	148	370	128	414
336	174	420	66	243	273	31	509
116	402	160	382	463	45	291	193
486	8	266	236	89	443	181	343
218	316	54	472	357	135	393	107
185	347	85	439	262	232	490	12
389	103	361	139	58	476	214	312

2

134	360	106	396	313	219	469	55
442	92	342	184	5	487	233	267
473	59	309	215	102	392	138	364
229	263	9	491	346	188	438	88
371	145	415	125	208	302	36	450
79	429	163	321	500	18	288	254
48	462	196	290	403	113	383	157
276	242	512	30	175	333	67	417

3

306	212	478	64	141	367	97	387
14	496	226	260	433	83	349	191
109	399	129	355	466	52	318	224
337	179	445	95	238	272	2	484
199	293	43	457	380	154	408	118
507	25	279	245	72	422	172	330
412	122	376	150	39	453	203	297
168	326	76	426	283	249	503	21

4

423	69	331	169	28	506	248	278
155	377	119	405	296	198	460	42
252	282	24	502	327	165	427	73
456	38	300	202	123	409	151	373
82	436	190	352	493	15	257	227
366	144	386	100	209	307	61	479
269	239	481	3	178	340	94	448
49	467	221	319	398	112	354	132

5

381	159	401	115	194	292	46	464
65	419	173	335	510	32	274	244
34	452	206	304	413	127	369	147
286	256	498	20	161	323	7	431
140	362	104	390	311	213	475	57
440	86	348	186	11	489	231	261
471	53	315	217	108	394	136	358
235	265	7	485	344	182	444	90

6

492	10	264	230	87	437	187	345
216	310	60	474	363	137	391	101
183	341	91	441	268	234	488	6
395	105	359	133	56	470	220	314
29	511	241	275	418	68	334	176
289	195	461	47	158	384	114	404
322	164	430	80	253	287	17	499
126	416	146	372	449	35	301	207

7

96	446	180	338	483	1	271	237
356	130	400	110	223	317	51	465
259	225	495	13	192	350	84	434
63	477	211	305	388	98	368	142
425	75	325	167	22	504	250	284
149	375	121	411	298	204	454	40
246	280	26	508	329	171	421	71
458	44	294	200	117	407	153	379

8

201	299	37	455	374	152	410	124
501	23	281	251	74	428	166	328
406	120	378	156	41	459	197	295
170	332	70	424	277	247	505	27
320	222	468	50	131	353	111	397
4	482	240	270	447	93	339	177
99	385	143	365	480	62	308	210
351	189	435	81	228	258	16	494

Sezione trasversale del cubo magico di ordine 8 di Richard Lewis Myers. (Tabulato gentilmente fornito da William Gosper.)

ECOLOGIA

LE SCIENZE
edizione italiana di
SCIENTIFIC AMERICAN

ha finora pubblicato su questo argomento i seguenti articoli:

CIRCOLAZIONE GLOBALE DELL'INQUINAMENTO ATMOSFERICO
di R.E. Newell (n. 32)

INQUINAMENTO DA MERCURIO
di L.J. Goldwater (n. 36)

L'ECOSISTEMA DEL PARCO DI SERENGETI
di R.H.V. Bell (n. 38)

I MODELLI MATEMATICI E L'AMBIENTE NATURALE
di R. Pennacchi (n. 45)

I CRATERI DELL'INDOCINA
di A.H. Westing e E.W. Pfeiffer (n. 48)

ENERGIA « PULITA » DA COMBUSTIBILI « SPORCHI »
di A.M. Squires (n. 53)

IL GRANDE DIBATTITO SUL BANDO AGLI ESPERIMENTI NUCLEARI
di H.F. York (n. 54)

IL CONTROLLO DEL CICLO DELL'ACQUA
di J.P. Peixoto e M. Ali Kettani (n. 59)

LA CRISI DELL'ACCIUGA PERUVIANA
di C.P. Idyll (n. 62)

LA FORESTA PLUVIALE TROPICALE
di P.W. Richards (n. 67)

L'OCEANO AL CONFINE CON L'ATMOSFERA
di F. MacIntyre (n. 72)

L'ELIMINAZIONE DEI RIFIUTI NELL'OCEANO
di W. Bascom (n. 76)

	3	4	5	6	7	8	9	10	11	12
3	2,2	*	2,3	*	2,4	2,5	2,5	*	2,6	3,3
4		3,3	*	3,4	4,1	*	*	3,8		
5			3,3	4,1	3,4	3,5	3,5			
6				4,4						

I risultati ottenuti da John Beidler a proposito del gioco di triple di Stanislaw Ulam.

tettivo e gli risposi suggerendogli l'indirizzo di una rivista matematica dove il suo lavoro sarebbe stato valutato.

In seguito udii parlare dei cubi di Myers nel dicembre del 1972 da John H. Staib, un matematico della Drexel University di Philadelphia, dove Myers si era iscritto proprio allora. Staib mi spedi un cubo di ordine 8 (*si veda l'illustrazione della pagina precedente*) e sebbene mettesse in luce le simmetrie di questa costruzione e spiegasse il metodo seguito da Myers (in cui si sovrappongono tre cubi latini e si fa uso della numerazione in base 8), continuai a non rendermi conto dell'importanza di questo cubo. Avevo infatti visto in maniera troppo superficiale il libro di Andrews, in cui si parla di cubi di ordine maggiore o uguale a 3, per rendermi conto del fatto che si trattava solo di cubi semiperfetti. È stato solo quando ho iniziato a lavorare a questo articolo che mi sono reso conto dell'importanza della realizzazione di Myers.

Dopo un anno di università Myers ha dovuto interrompere gli studi e oggi lavora alla Computaprint Corporation a Fort Washington, Pennsylvania, come programmatore, dove spera di guadagnare sufficiente denaro per riprendere gli studi di matematica.

Ogni linea ortogonale o diagonale di otto numeri nel cubo di Myers mostrato nella pagina precedente, incluse le quattro diagonali spaziali, dà come somma 2052. Il cubo è associativo: due numeri opposti simmetricamente rispetto al centro danno come somma 513. Ne segue che non solo gli otto numeri d'angolo danno come somma 2052, ma anche i numeri corrispondenti agli angoli di qualsiasi solido rettangolare inscritto nel cubo danno la stessa costante. Ma non è tutto: è possibile suddividere il cubo in 64 cubetti di ordine 2 in modo che gli otto numeri di ognuno dei cubetti diano come somma la costante!

L'esistenza di queste notevoli simmetrie ha reso possibili molte varianti, tutte in un certo senso isomorfe, del medesimo cubo, a cui bisogna naturalmente aggiungere, per ognuna di esse, le 48 possibilità derivanti da rotazioni e da riflessioni. Si provi a immaginare questo cubo con ognuna delle sue 512 celle rimpiazzata dal cubo stesso in una qualsiasi delle sue varianti. Nella cella 1 mettiamo un cubo che inizi con 1, nella cella 2 mettiamo un cubo che inizi con $8^3 + 1 = 513$, nella cella tre un cubo che inizi con $(2 \times 8^3) + 1 = 1025$ e così via. Il risultato è un cubo magico perfetto di ordine 64. A partire dall'ordine 64 si può costruire con lo stesso procedimento un cubo magico perfetto di ordine 512, e così via per tutte le potenze di 8.

Quanti sono i cubi magici perfetti di ordine 8? Scegliendo differenti cubi latini da sovrapporre, Myers può costruire milioni di cubi diversi l'uno dall'altro e da quello qui presentato, anche se non tutti associativi. I quadrati latini di ordine 8 sono stati enumerati (ve ne sono miliardi), ma non i cubi latini: ciò dà l'idea della difficoltà di enumerare i cubi di ordine 8 generabili con il metodo di Myers.

È 8 l'ordine minimo per avere un cubo magico perfetto? Vi sono cubi magici perfetti che non rientrano nelle potenze di 8? Sono ancora problemi insoluti.

Per quanto riguarda i problemi del mese scorso, quello riguardante le frazioni composte in cui si richiedeva di trovare quattro numeri a, b, c, d per cui valesse $(a/b)/(c/d) = (d/c)(b/a)$, era evidentemente uno scherzo, dato che l'espressione è una identità considerando numeri reali.

La soluzione del criptaritmo di Alan Wayne, SIX + SIX + SIX = NINE + NINE, è 942 + 942 + 942 = 1413 + 1413. Si noti che 1413 sono le prime quattro cifre di π lette alla rovescia, e che 942/3 dà 314, le prime tre cifre di π.

Nel fascicolo di gennaio è stato presentato un gioco di Stanislaw Ulam, simile al cram giocato coi tromino. John Beidler, direttore del dipartimento di teoria dei calcolatori dell'Università di Scranton, ha trovato che giocando su una scacchiera standard di 6 per 6, il primo giocatore ha la vittoria solo se muove su una delle quattro caselle centrali. Beidler ha generalizzato il gioco al caso di scacchiere rettangolari, ottenendo il risultato mostrato nella figura in alto di questa pagina. I numeri danno le mosse vincenti del primo giocatore indicando la riga e la colonna. Gli asterischi indicano i casi in cui la vittoria tocca al secondo giocatore.